THE UNIVERSITY OF CHICAGO

STATISTICAL METHODS FOR UNRAVELING REGULATORY GENOMICS

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES

AND

THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES

AND THE PRITZKER SCHOOL OF MEDICINE

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN BIOPHYSICAL SCIENCES

BY

YIFAN ZHOU

CHICAGO, ILLINOIS

JUNE 2022

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to start by thanking the people who supported me in my scientific endeavors. In particular, I would like to acknowledge my advisors, Professor Xin He and Assistant Professor Mengjie Chen, who have constantly been patient with me and supported me throughout my PhD. Xin played an essential role in my PhD as my research advisor and my scientific role model. I am grateful for the enormous effort he dedicated into my scientific development. When I first joined Xin's group, we worked together every week to fill in the gaps of my knowledge and ultimately give me scientific foundation in the field; later on, we still met to either learn about the latest developments in the field, or trouble-shoot specific problems. To this day, I have yet to witness another advisor as patient and invested in the continuous learning of their students as Xin. I am grateful to have been a part of his group, where he constantly leads by example with his intellectual curiosity, scientific rigor and big-picture thinking.

My other research advisor, Mengjie, was also instrumental in my scientific and personal development. With her sharp mind and extensive knowledge in applied statistics, Mengjie was always ready to help when I needed it. Her assertive, practical, and punctual advice was critical for converting abstract statistical ideas into intuitive, pragmatic methods in a data-driven manner. I am also grateful for Mengjie for the encouragement she offered whenever I had confusion about my personal trajectory.

I also want to express my gratitude to my committee members, Professor Matthew Stephens, Professor Marcelo Nobrega, and Professor Ivan Moskowitz, for their kind and helpful guidance during the course of my research. It has indeed been a privilege, being able to work on the "fourth floor" of Cummings Life Science Center surrounded by so many brilliant minds in the field of

human genetics. I am particularly grateful for the helpful insights from Matthew, who could always grasp the essence of scientific problems, and inspire new ideas in us to tackle the obstacles at hand.

Finally, I would like to thank the people closest to me. I am eternally thankful to my parents, who have nurtured the scientific curiosity in me since an early age, and have provided unconditional love and support throughout my endeavors in life. I also want to thank my boyfriend, Alan, who I met in the beginning of my PhD, have grown very close to in the end. The scientific discussions we had and the emotional support he provided made the journey less difficult.

ABSTRACT

The genetic code carries instructions for the development and functioning of every biological organism. Variation in this code may cause mis-regulation of genes expression, affect cellular states, and ultimately lead to observable changes in organism-level traits. Genome Wide Association Studies (GWAS) have discovered thousands of significant statistical associations between single-nucleotide polymorphisms (SNPs) and disease/traits in human, however, functional interpretation of these associations remains challenging. To gain mechanistic insights into the relationship between genetic variations and their phenotypes, a comprehensive understanding of the gene regulatory architecture is the first and fundamental step. This dissertation addresses some of the challenges in unravelling the regulatory functions in non-coding regions and effects of genetic variations at the transcription level. I develop novel computational frameworks and statistical methods that complement experimental approaches, which combined together, aid the discovery of regulatory elements and functional disease variants, and improve the understanding of the genetic basis of diseases. In Chapter 1, using ATAC-seq and RNA-seq data of human neurons, I map *cis*-regulatory elements and investigate their interactions with transcription factors and target genes, deriving preliminary gene regulatory networks for autism risk genes. In Chapter 2, we outline a novel framework for integrating GWAS results with the allelic-imbalance open chromatin (ASoC) information captured by ATAC-seq. Leveraging ASoC in neurons, we prioritize putative causal non-coding SNP in schizophrenia GWAS. Data analysis of single-cell CRISPRi screen with RNA-seq readout further confirm the regulatory functions at six SNP loci and their corresponding target genes. However, due to the novelty of high-throughput single-cell CRISPR screen technologies, statistical methods for effective analysis and interpretation of such data are lacking. In Chapter 3, I develop a novel Bayesian factor analysis

method that can detect from these perturbed expression data genes and gene modules impacted by the CRISPR perturbations. I apply this method to simulated and publicly available datasets. In addition to identifying biologically relevant gene modules, the method has better power to detect differentially expressed genes than alternative methods, shedding light on the regulatory basis underlying T cell activation and neuronal differentiation.

INTRODUCTION

The genetic code carries instructions for the development and functioning of every biological organism. For humans, the genetic code lies in the human genome that consists of 3 billion DNA base pairs consisting of four nucleotides – A, T, C, G. Variation in this code may alter the expression of genes and other molecular intermediates, disrupt cellular processes, and ultimately lead to observable changes in human traits. Therefore, it is crucial to understand the functions of each part of the genetic code to our best ability, in order to unravel the effective genetic variants and their causal mechanisms for any disease or trait.

The first effort to map the human genome was completed in 2003, revealing that, out of the 3 billion base pairs, only < 2% are used to encode proteins (*i.e.* coding sequences), while the rest 98% are non-coding sequences that are poorly understood[1]. Almost a decade later, more and more evidence suggests that these non-coding regions are rich in elements that precisely control the spatiotemporal expression of genes[2], a process that is far more complex than previously thought. To this date, many questions remain unknown about the genome in terms of gene regulation: Which and how regulatory elements are involved? Do genetic variations cause mis-regulation and how are they relevant to observed disease phenotypes?

Gene expression control and regulatory elements

The precise control of gene expression at the transcription level involves the coordination among regulatory sequences and proteins. *Cis*-regulatory elements (CREs) such as promoters and enhancers reside in the non-coding genome, and contain binding sites for regulatory proteins – transcription factors (TFs). Working together, they can either recruit or block RNA polymerase II at a gene's promoter, thereby modulating its transcription level. These regulatory components fine-

tune the intricate transcriptional programs in each cell type, maintaining cellular functions and processes that are either general or cell-type-specific[2].

Chromatin exists as a dynamic structure in the nucleus, with its compactness level modulated to occlude or allow the access of chromatin-binding factors to DNA[3]. Since CREs need to be available for TF binding, these functional sequences are often marked by accessible chromatin. The accessibility or compactness of chromatin can be measured by DNase I hypersensitive sites sequencing (DNase-seq)[4], and most recently, Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq)[5], both of which use cleavage enzymes (DNase-I and Tn5 transposase, respectively) to extract and sequence the DNA wherever they can access. Although these technologies can predict the location of CREs across genome, not all accessible regions correspond to active CREs. For example, an enhancer might be inactive but primed open for activation at a later time point, or it might be accessible in multiple cell types but only active in a cell-type-specific manner[6]. Additional information such as the modification of histone proteins can be utilized to identify active CREs marked by H3K27ac[7]. Active CREs also tend to have bound TFs. TF binding sites along the DNA can be directly sequenced and identified through chromatin immunoprecipitation sequencing (ChIP-seq)[8]. Careful computational analyses on high-resolution chromatin accessibility data can also reveal TF binding footprints within the accessibility signal where bound TFs protect their underlying DNA from cleavage[9], indirectly informing TF binding events.

Besides genome-wide prediction of CRE location through chromatin features and TF binding, it is also important to understand what their regulatory targets are. However, it is now understood that CREs can act either locally or over long distances via chromatin looping[10] to modulate the expression of their target genes, without any specific patterns in their range, rendering the

identification of their target genes challenging. Several recent developments (3C-based technologies[10], [11]) have enabled the measurement of physical interactions within chromatin in the nucleus, a particularly sensitive one being promoter capture Hi-C[12], which can sequence and identify genomic regions that have physical interactions with gene promoters, directly linking putative distal CREs to the target genes. However, it is important to keep in mind that 3C-based methods only assess physical contacts and do not necessarily reflect regulatory relationships; additional investigations are needed to determine causal regulatory relationships between regions in contact with each other[6].

Study of genetic variations

Variability in the genetic code has been a focus of human genetic research as it leads to different gene expression levels and other downstream phenotypes, which results in varying heritable disease risks among individuals. The most common type of genome variation among the population are single nucleotide polymorphisms (SNPs). The relationship between SNPs and thousands of human traits/diseases have been extensively studied in Genome Wide Association Studies (GWAS), where the presence of a SNP is statistically associated with whether an individual possesses a certain trait over the genotyping of large numbers of individuals[13]. While GWAS has identified numerous genomic variants associated with complex disease and tremendously aided our understanding of their genetic bases, interpretation of these associations remains challenging. Most SNPs found significantly associated with a disease are not causal, but rather tag SNPs merely chosen because they are in high linkage disequilibrium (LD) with the actual causal SNP. Therefore, it is difficult to pinpoint the disease-driving variants and the genes these variants act through. Another way to approach this is by examining the genomic context of these significant associations for the ones with important functions. Since over 90% of SNPs associated

with complex traits are mapped to the non-coding genome[14], the characterization of regulatory elements and their target genes are of great importance. Efforts are needed to understand the regulatory effects of genetic variants, as it will lead to identification of causal variants and facilitate a mechanistic understanding of the genetic basis of diseases.

While we have introduced many ways to characterize CREs and the genes they interact with, they do not reflect causal regulatory relationships. One way to establish the causal link is to directly perturb the genomic sequence of interest and evaluate the phenotypic outcomes of the perturbation in cells. CRISPR genome editing[15] has enabled the perturbation of genomic sequences in their endogenous context, providing researchers with a powerful tool to screen for CREs and study their functions[16]. CRISPR coupled with the Cas9 nuclease can induce a double-strand break at the targeted regulatory element under the direction of a guide RNA (gRNA), resulting in disruptive genetic effects sometimes comparable to the deletion of the whole enhancer[17]. This system can also be modified to achieve modulation of enhancer activity using a nuclease-deactivated Cas9 (dCas9) instead. For example, sequences can be repressed through CRISPRi, where dCas9 is fused with repressive effectors such as the KRAB domain[18]. By designing gRNAs that target different genomic loci and delivering them into cells using lentiviral vectors, researchers can perform functional screens of multiple CREs in a single experiment[19]. Most recently, technologies such as Perturb-seq[20], [21] and CROP-seq[22] have further improved the resolution and throughput of genetic screens by combining pooled CRISPR screening with single-cell RNA-seq (scRNA-seq), allowing for efficient screening for effects of multiple genetic perturbations in tens of thousands of cells simultaneously[23].

In general, single-cell technologies have demonstrated substantial advantages over bulk assays in their ability to capture cellular heterogeneity in transcriptomic and epigenetic programs.

When applied to tissues of multiple cell types, or dynamic systems such as cells undergoing differentiation, these technologies can tremendously aid the discovery of context-dependent or cell-type-specific regulatory relationships[24], [25]. However, several challenges exist in the statistical analyses of single-cell CRIPSR screen data. First and foremost, the sparse and noisy nature of raw single-cell genomic data introduces difficulty with normalization[26] and violates the asymptotic assumptions of parametric statistical tests if modeled improperly. A solution that has been adopted by the field is to perform exact tests either through permutations[27], [28], which are computationally intensive, or utilizing non-parametric approaches such as the Wilcoxon rank sum test. In addition, because of low gRNA efficiency, relatively small numbers of cells are assigned to each gRNA-mediated perturbation, further limiting the discovery power of these experiments. As a result, routine differential expression analyses often only manage to identify a fraction of effects, and the results vary drastically from method to method[29]. Better data-driven methods are needed to increase the detection power of effects of genetic perturbations in single-cell CRIPSR screen data.

Overview of thesis research

In this thesis, I attempted to address some of the challenges in understanding regulatory regions and effects of genetic perturbations, and developed novel computational frameworks and statistical methods that complement experimental approaches, which combined together, aided the discovery of regulatory elements and functional disease variants, and improved the understanding of the genetic basis of diseases.

Chapter 1 is a preliminary effort to understand the gene regulatory structure in developing human neurons. Integrating matched bulk ATAC-seq and RNA-seq data from human induced pluripotent stem cells (iPSCs) and derived neurons, I mapped genome-wide putative CREs and

predicted TF binding events in these CREs from chromatin accessibility data; then, by modeling the relationship between gene expression and CRE accessibility, I identified CREs that are highly associated with genes of interest. Under this framework, I established a preliminary gene regulatory network that connects the *trans*-regulatory factors to the *cis*-regulatory elements, and to their potential target genes for a set of autism risk genes.

While these mapped CREs, or more generally, open chromatin regions, are more likely to have regulatory potential, not all disease risk variants located in them are functional. As a result, the pinpointing of causal functional variants remains challenging for complex traits such as neuropsychiatric disorders. In Chapter 2, we provided a more comprehensive framework to address this challenge using the same study system. We focused on variants in the genome that exhibit allelic-specific open chromatin (ASoC). We believe that these ASoC variants are more likely to be functional and disease-relevant based on the simple fact that changes in accessibility can impact gene expression, which is one of the main disease-causing mechanisms. Characterization of mapped ASoC variants confirmed their greater tendency to be functional than other regular variants. For example, neuronal ASoC variants were found enriched for brain enhancers, TF binding sites, and brain QTLs. After confirming their high disease relevance using GWAS schizophrenia (SZ) summary statistics, we prioritized a set of GWAS SZ SNPs that are also neuronal ASoC variants, and further validated on their regulatory functions using CRISPR genome editing in both single cells and bulk cell lines. Overall, we identified 6 SZ risk variant loci and 9 corresponding *cis*-target genes regulated by them. These findings demonstrated the effectiveness of ASoC in interpreting the effects of non-coding disease variants in a neurodevelopment context, which can aid the genetic discovery of brain-related traits.

While analyzing single-cell RNA-seq with CRISPR screen (CROP-seq) data in Chapter 2, we found that routine differential expression (DE) methods were under-powered in detecting the transcriptome effects of genetic perturbations. In Chapter 3, we focused on addressing this specific computational challenge, and developed a novel statistical method to better analyze single-cell RNA-seq with CRISPR screen data. Instead of interrogating the change in expression for each gene individually, our factor analysis-based model leverages the modular structure in the transcriptome, and detects coordinated modules of genes that are affected by the perturbations. In addition to the detection of affected gene modules, our method also summarizes the effect of perturbation for each gene, allowing for the discovery of differentially expressed genes. Applying our method to two external CROP-seq datasets, we demonstrated that it can uncover biologically relevant gene modules and has better power to detect differentially expressed genes than routine DE methods, which together, improved the understanding of regulatory mechanisms underlying T cell activation and neuronal differentiation.

CHAPTER 1: INFERENCE OF GENE REGULATORY NETWORKS AT THE TRANSCRIPTION LEVEL

1.1 Introduction

Many regulatory components are involved in the precise control of gene transcription, forming a complex network. Deciphering gene regulatory architectures is of great interest, as it would highlight functional regions of the non-coding genome, unravel the genes they mediate their effects through, and help explain the varying heritable disease risks across individuals[16].

One major component of the gene regulatory network is *cis*-regulatory elements (CREs), the most understood of them being promoters and enhancers. CREs typically contain the binding sites for transcription factors (TFs), and the coordination of these regulatory elements are necessary to activate or repress the transcription of genes. CREs can be characterized by assays of chromatin accessibility such as DNase-seq[4] and ATAC-seq[5]. However, an important discovery of enhancers is that they can act independently of the distance and orientation to their target genes[30], and function over great distances (up to 1 Mb) through DNA looping[10] to modulate the expression of a gene at its promoter, which makes the identification of their target genes difficult. Chromatin conformation capture based technologies such as Hi-C[12] can provide snapshots of physical interactions within chromosomes in the nucleus, enabling the mapping of long-range interactions between enhancers and promoters. Yet, it is still challenging for these assays to provide enough spatial resolution to identify specific contacts between enhancers and gene promoters; in addition, because these interactions are dynamic and often cell-type- or tissue-specific, they may not be fully captured by these assays. As for TFs, although their binding to the genome can be directly measured through ChIP-seq[8], these assays are low-throughput; and given the prevalence of non-functional binding events in the genome, a large portion of TF binding sites

detected by ChIP-seq have no regulatory roles[31]. Therefore, data from various orthogonal sources need to be integrated to establish regulatory relationships among TFs, CREs, and their target genes for the specific cell types of interest.

Here, we utilized a series of statistical approaches to complement the experimental limitations and constructed cell-type specific gene regulatory networks in a neurodevelopmental model consists of human induced pluripotent stem cells (iPSCs) and derived neurons, leveraging matched bulk ATAC-seq and RNA-seq data. We mapped cell-type-specific putative CRE regions based on ATAC-seq data. Applying TF footprint calling to the ATAC-seq data, we established the connection between TFs and putative CREs. By modeling the relationship between CRE activity and gene expression, we quantified the link between CREs and target genes. Collectively, we constructed preliminary gene regulatory networks of several high-confidence autism risk genes, which not only serves as a resource for regulatory components potentially involved in the modulation of these genes, but may also help in predicting the effects of perturbations in CREs on gene expression. We envision that application of these findings on expression quantitative trait loci (eQTL) analyses or genome-wide association studies (GWAS) will help elucidate the genotype-phenotype associations and prioritize causal SNPs for crucial brain disorders.

1.2 Methods

1.2.1 Chromatin accessibility profiling using ATAC-seq[1]

8 reprogrammed iPSC lines were obtained from 8 subjects in the Molecular Genetics of Schizophrenia (MGS) cohort, and differentiated into 4 neuronal cell types: neural progenitor cell (NPC), glutamatergic neuron (iN-Glut), dopaminergic neuron (iN-DA), and GABAergic neuron

[1] Sections 1.2.1 and 1.3.1 contain material from the paper: Zhang *et al.*, "Allele-specific open chromatin in human iPSC neurons elucidates functional disease variants," *Science*, vol. 369, no. 6503, pp. 561–565, 2020, doi.org/10.1126/science.aay3983.

(iN-GA) (Fig. 1.1a). Bulk ATAC-seq and RNA-seq were performed on these 40 samples individually.

The paired-end and single-end ATAC-seq reads of each sample were mapped to GRCh38p7 using Bowtie[32], then merged and sorted as BAM-formatted files using SAMtools[33]. Processed BAM ATAC-seq reads from samples of the same cell type were pooled together to increase the read depth, and therefore, the power of downstream peak detection. We used MACS2 to generate piled-up coverage of ATAC-seq cut-sites and perform peak calling with settings --nomodel, --shift -100, --extsize 200, --slocal 1000, --llocal 10000, and --qvalue 0.01. Peaks called from all cell types were consolidated to form a union set of peaks. The chromosomal locations of these peaks correspond to putative CREs, and reads that overlap with these intervals were re-counted for each sample using bedtools, reflecting sample-specific CRE activities. A Snakemake pipeline that processes ATAC-seq bam files of individual samples and converts them into a structured peak count matrix as described is available at github.com/gradonion/ATACseq_pipeline. Conditional quantile normalization[34] was applied to the raw peak count matrix to account for the differences in library size, peak length and GC content using R package CQN. Unless stated otherwise, this normalized peak activity matrix was used for all downstream analyses.

Bulk RNA-seq data for each sample was mapped to GRCh38p7 using STAR[35]; duplicates were removed using Picard MarkDuplicates, and gene counts were generated using featureCounts. Similarly, conditional quantile normalization was applied to the raw gene count matrix to account for gene length, GC content, and sample library size. Unless stated otherwise, this normalized gene expression matrix was used for all downstream analyses.

Principal component analysis (PCA) and hierarchical clustering were performed on both the peak activity profiles and gene expression profiles to examine the structure within samples. For gene expression, only the top 10k most variable genes were used in the analyses.

1.2.2 Transcription factor footprint calling

DNase-seq and ATAC-seq, two assays that are widely used for genome wide profiling of chromatin accessibility, share similar mechanism in that they both use cleavage enzymes (DNase-I in DNase-seq[4] and Tn5 transposase in ATAC-seq[5]) to cleave the accessible parts of the DNA. Interestingly, DNA accessibility data obtained using this mechanism also inform on transcription factor binding and nucleosome positioning. As TF binding protects the DNA from enzyme cleavage, this leaves a local window that drops in accessibility coverage, referred to as a TF footprint[36]. This information can be used to complement the traditional motif analysis approach, where regions matching the known TF PWMs are identified purely based on their sequences. By recognizing motif instances that also exhibit depletion of DNase-seq or ATAC-seq signals locally, functional TF binding sites (TFBSs) can be predicted with more confidence.

Several computational tools were built upon this idea to predict TFBSs. One approach, CENTIPEDE[37], utilizes the difference in accessibility profiles between bound and unbound TF motif instances, and trains a Bayesian mixture model on the piled-up cut counts surrounding all motif instances in the genome for the TF motif of interest. Motif occurrences predicted to be in the binding states are then considered as TFBSs. This unsupervised method does not make assumptions on the depletion of cut counts or set hard thresholds to distinguish bound and unbound states, and therefore, can identify TFBSs with high sensitivity. However, numerous preprocessing steps are needed to prepare an appropriate input for the CENTIPEDE software, including the scanning of genome-wide motif instances and the extraction of DNase-seq or ATAC-seq count

profiles in a 200bp window centered on each of the motif matches. This adds substantial complexity and processing time to the approach. In our case, the prediction of TFBSs for a single TF motif takes > 1 hour starting from a processed ATAC-seq BAM file pooled over a cell type.

Most TF footprinting methods were initially developed for DNase-seq. Although similar in principle, the transposase Tn5 used in ATAC-seq has a different cleavage bias than DNase-I[38]. HINT-ATAC[38] is the first footprinting method to account for such bias in ATAC-seq data. It first uses position-dependent models to estimate the cleavage bias of Tn5 from a given ATAC-seq library; then, using the corrected cleavage signals as inputs, it trains an HMM model to infer the states of motif instances. Prediction for 579 TF motifs takes ~3.5hr. Although HINT-ATAC was reported to out-perform other competitors in ATAC-seq footprint prediction[38], we did not observe the same trend in our study (see below).

Here, we adopted a two-step footprint calling approach (Fig. 1.1c), where we first used DNase2TF[36] to search for putative footprint windows of within ATAC-seq peak regions depleted of cleavage events, and then overlapped them with high-confidence TF motif matches in the OCRs found by a motif analysis tool ('rgt-motifanalysis') in Regulatory Genomics Toolbox (RGT) to obtain TFBSs for 579 curated TF motifs in the JASPAR Core database (2018 release). Both steps are fast in computation (all 579 motifs in under 20 minutes) and can be carried out independently of one another.

DNase2TF[36] relies on a relatively simple algorithm that directly scans along the accessibility signals for depletion regions. It starts with a set of candidate regions of sizes ranging from 6 bp to 30 bp within the OCRs that are local minima in the cut count profile, and assesses the significance of count depletion at each candidate region by comparing against a wider local background region centered at the candidate region using a binomial test. In the analysis, the size

of a local background region is fixed to be k (=3) times the size of the candidate region. Assuming that sequencing reads are distributed uniformly across the background and N cut counts are observed in the local background region, the number of cut counts in the candidate region should follow the binomial distribution Binom(N, 1/k). This distribution can be approximated by a normal distribution when N is large, generating the following z score for a candidate region with n observed cut counts:

$$z = \frac{n - \frac{N}{k}}{\sqrt{\frac{N}{k}(1 - \frac{1}{k})}}.$$

The lower the z score is, the more depleted the candidate region is of ATAC-seq cleavage events. A false discovery rate (FDR) is further estimated for each candidate by repeating the algorithm for randomized cut count data. Finally, a set of candidate footprint regions can be obtained by thresholding the z score (< -2) and FDR (< 0.01). Although additional analysis of TF motif matching is necessary to predict TFBSs, DNase2TF holds advantages over other TF footprint calling methods of its time in its fast computation and relatively high prediction accuracy[36].

Our second step is to use a motif matching tool 'rgt-motifanalysis'[39] to identify TF motif instances in the ATAC-seq peaks. The tool evaluates the match between every possible sequence and a given TF position weight matrix with a log-likelihood ratio score, then generates a false positive rate (FPR) based on the score using dynamic programming, and finally accepts motif instances by FPR thresholding. For each JASPAR TF motif, we obtained matching motif instances within the ATAC-seq peaks under an FPR threshold of 10^{-4}. Among them, instances that overlapped by at least 50% with the potential footprint windows identified by DNase2TF were determined as the final TFBSs. This procedure of TFBS prediction, or TF footprint calling, was

performed for each cell type, where ATAC-seq cut counts from samples of the same cell type were pooled together.

We evaluated the performance of TF footprint calling using ENCODE TF ChIP-seq data measured in embryonic stem cell lines (H1-hESC), which is the closest cell type to iPSC, as the ground truth. H1-hESC ChIP-seq TFBS uniform peaks of CTCF, GABPA, SP1, and ZNF143 were downloaded from https://genome.ucsc.edu/ENCODE/dataMatrix/encodeChipMatrixHuman.html in .narrowPeak format, converted to hg38 using liftover, and compared with our predicted TFBSs of corresponding TFs in iPSC. To assess the sensitivity and specificity of our footprint calling method for a specific TF, we treated all its matching motif instances in the OCRs as the testing pool, and defined true positives as motif instances that are called as footprints and have ChIP-seq signals, while false positives are defined as motif instances that are called as footprints but do not have ChIP-seq signals (Fig. 1.1d). ROC curves for our approach were generated by varying the FDR threshold in DNase2TF.

While we did not generate ROC curves for CENTIPEDE and HINT-ATAC, we evaluated their performance in predicting CTCF TFBSs in iPSC at realistic thresholds that resulted in comparable number of discoveries. Using a posterior probability threshold of 0.99, CENTIPEDE predicted 1.83×10^4 CTCF footprints in iPSC with a precision of 0.92, while HINT-ATAC predicted 1.24×10^4 CTCF footprints with a precision of 0.59 under the default setting. Our approach involving DNase2TF performed between these two methods, detecting 1.77×10^4 CTCF footprints with a precision of 0.84 (DNase2TF FDR < 0.01). Evaluation on the other 3 TF motifs yielded similar results, with the performance of our approach ranking in between CENTIPEDE and HINT-ATAC, which justified our use of DNase2TF since it is on average ~100 times faster than CENTIPEDE.

1.2.3 Statistical modeling of the relationship between CRE activity and gene expression

Human iPSC-derived neurons are a powerful neurodevelopmental model that can be used to study related disorders such as autism spectrum disorder (ASD) and schizophrenia[40], [41]. Here, we set out to understand the regulatory components of ASD risk genes. Except for promoters that locate immediately upstream of the TSS of genes they regulate, there is no clear pattern of the relative location of CREs to their target genes, and since high-resolution Hi-C data for matching cell types were unavailable, we decided to infer these functional links by statistically modeling the relationship between CRE activity and gene expression.

We chose a set of 86 high-confidence ASD-associated genes with a score of 1 or 2 in the SFARI database[42] as target genes. For a given target gene, we consider ATAC-seq peaks within 100 kb region upstream of its TSS or introns as the candidate set of regulatory elements that potentially regulate the expression of the target gene. We proposed a prediction model where the linear combination of the activity levels of these peaks (measured by ATAC-seq) determines the expression level of the target gene (measured by RNA-seq):

$$y_i \sim \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_P x_{iP},$$

where y_i is the normalized (CQN) expression level of target gene G in sample i, and x_{ip} is the normalized (CQN) peak strength of the p-th candidate peak for gene G in sample i (i=1, 2, …, N). We excluded the proximal promoter from the prediction, as its function is often well-understood, while its activity tends to strongly correlate with gene expression, which can mask the minor effects of enhancers. In reality, given our definition of peak assignment, there can be up to hundreds of candidate peaks under consideration for a given target gene, while sample size N = 40. To ensure the interpretability and improve the accuracy of the regression model, variable

selection or regularization approaches are needed to select only a subset of peaks that have strong association with the gene expression outcome.

Here we adopted a regularization approach, elastic net linear regression[43], to select an appropriate number of associated peaks and obtain interpretable models. Another popular and successful approach that uses regularization penalties to achieve parsimonious models is lasso[44]. But since the number of predictors selected by lasso is bounded by the number of samples, and it tends to fail in selecting grouped variables, it is not ideal for this problem. Combining lasso and ridge regression penalties, elastic net overcomes the limitations of both methods, and thus is more suitable for handling large numbers of correlated predictors[43]. The estimates of effect sizes take the following form under our elastic net regularization:

$$\widehat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \left\{ \|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \frac{0.5\lambda}{2}\|\boldsymbol{\beta}\|^2 + 0.5\lambda\|\boldsymbol{\beta}\|_1 \right\},$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_P)$, $\boldsymbol{y} = (y_1, y_2, \ldots, y_N)$, $\mathbf{X}$ is the $N \times P$ matrix of x_{ip}'s, and λ is the regularization parameter that controls the amount of shrinkage on $\boldsymbol{\beta}$. In practice, the value of λ is determined by 10-fold cross validation for each target gene. Specifically, for each gene and its candidate peaks, we first used the cv.glmnet() function in the R package glmnet to evaluate the fit under different λ's, and then picked λ as the largest value such that the cross-validation error is within one standard error of the minimum to encourage sparsity. Finally, we fit our elastic net regression model under the chosen λ using the glmnet() function, obtaining a set of selected peaks that have non-zero association effect sizes, or in other words, are potential CREs of the target gene.

To understand what differentiates the selected peaks from the unselected candidate peaks, we further investigated the enrichment of several epigenetic features in these selected peaks. Suppose for a pool of C candidate peaks, each peak has one of the binary states represented by s_j, with 1 for being selected, and 0 for not being selected. And we can quantify a certain genomic feature (*e.g.*

the number of TFBSs) at each peak region as f_j. The enrichment level (α_1) of this genomic feature in selected peaks can then be estimated using logistic regression:

$$\log\left(\frac{p(s_j = 1)}{p(s_j = 0)}\right) = \alpha_1 f_j + \alpha_0, \qquad j = 1,2,\dots,C.$$

To avoid any biases introduced by different scaling across features, the genomic features we investigated are all binary, including (i) whether a peak contains predicted TFBSs, (ii) whether it is an intronic peak of the target gene, (iii) whether it is a nearby peak within 20kb upstream of TSS of the target gene, (iv) whether the peak contains active H3K27ac histone modification mark (fetal brain H3K27ac ChIP-seq data from[45]), and (v) the binary GERP (genomic evolutionary rate profiling)[46] score of sequence conservation at the peak region.

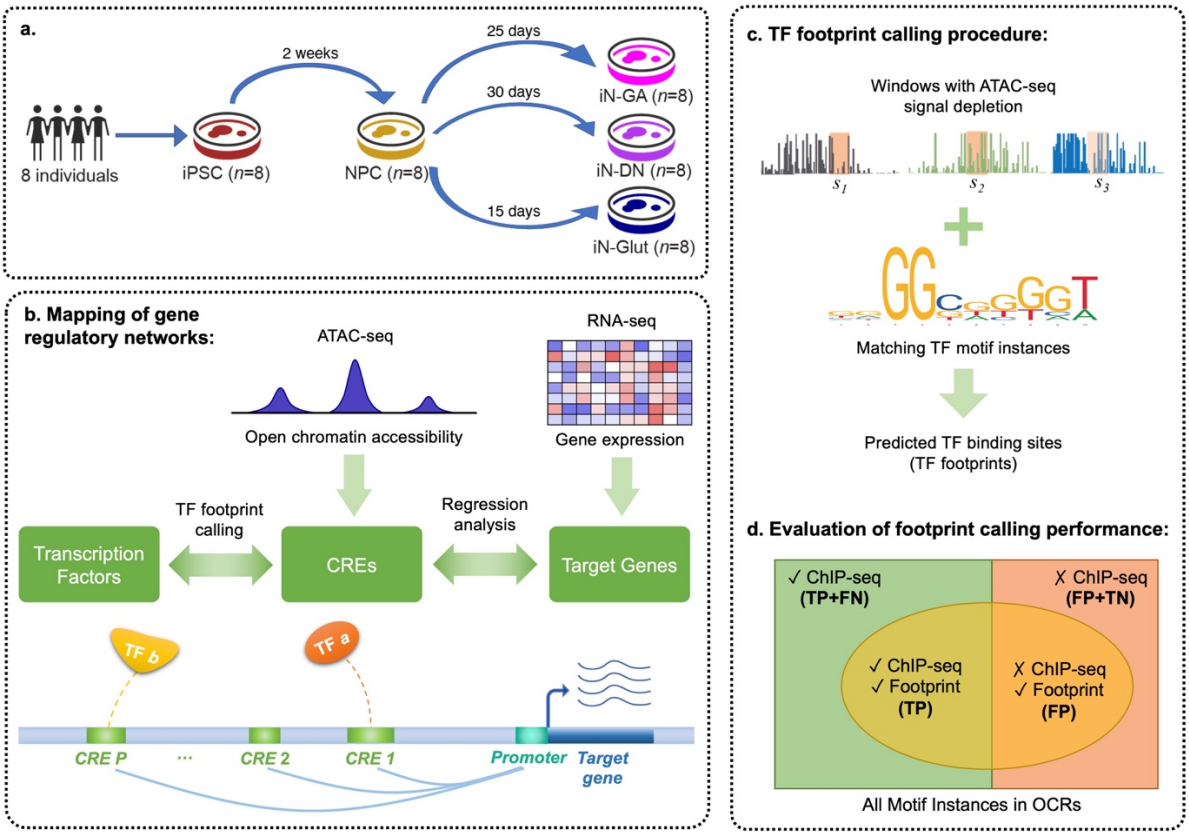

Figure 1.1 Schematic of experimental data and analysis workflow.

1.3 Results

1.3.1 Identification of cell-type-specific *cis*-regulatory elements (CREs)

To identify the OCRs with potential regulatory functions in our neurodevelopmental model, we performed ATAC-seq in each of the 40 cell line samples from 5 cell types (8 samples per cell type). We pooled the ATAC-seq reads from samples of the same cell type and conducted peak calling in the piled-up cut count coverages, identifying 302K peaks in iPSC, 152K peaks in NPC, 157K peaks in iN-Glut, 237K peaks in iN-DA, and 261K peaks in iN-GA, spanning 2.61% to 5.73% of the whole genome. Combined together, these peaks form a union set of 510K peaks in total across all cell types, 53.2% of which are cell-type-specific.

As expected, principal component analysis (PCA) and hierarchical clustering of sample ATAC-seq peak profiles both exhibited clustering of samples of the same cell type, with the neuronal samples being more similar to each other than to iPSC samples (Figs. 1.2a,b). Similar patterns were observed among sample RNA-seq profiles (Figs. 1.2c,d). Our ATAC-seq data also reflect cell-type-specific chromatin accessibility at cell-type-specific marker gene loci. For example, for NANOG, an important TF for stem cells to maintain their pluripotency, genomic regions near its gene locus are the most accessible in iPSCs than in neuronal cell types (Fig. 1.2e). *GAD2*, a gene primarily expressed in GABAergic neurons, shows higher accessibility in its TSS in iN-GA than in the rest neuronal cell types (Fig. 1.2f).

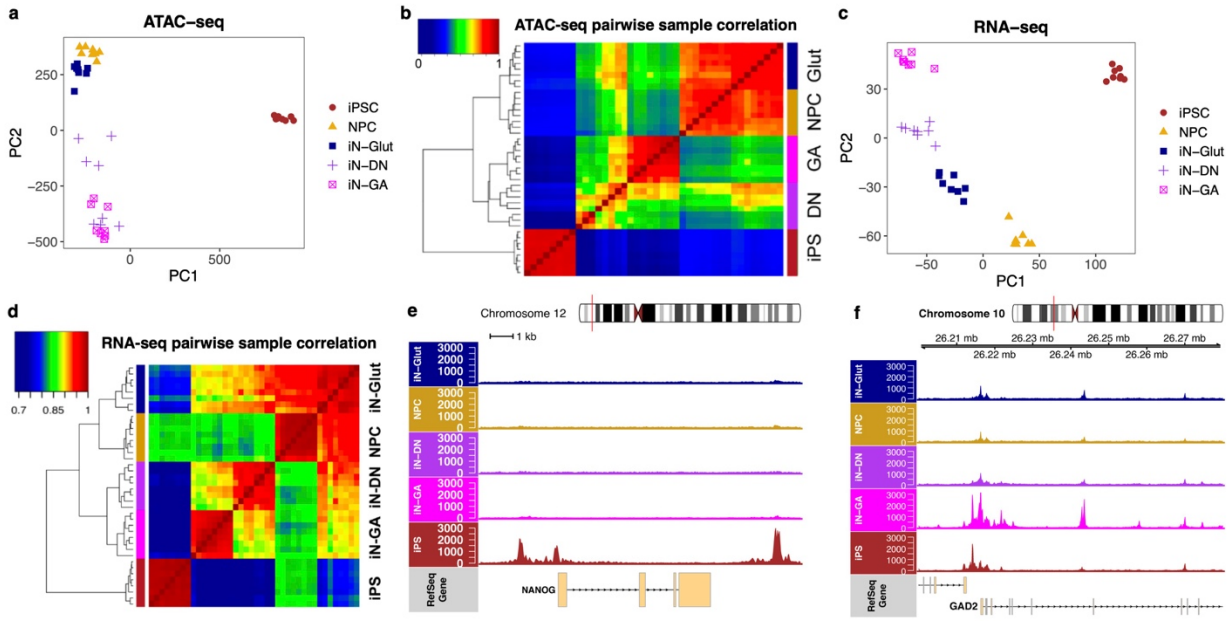

Figure 1.2: ATAC-seq open chromatin peak activity and RNA-seq gene expression profiles in different cell types.

a) PCA dimensional reduction of ATAC-seq peak count data of 40 samples from 5 cell types. **b**) Hierarchical clustering based on ATAC-seq peak count data of 40 samples, with heatmap colors reflecting the pairwise correlation between samples. **c**) PCA dimensional reduction of RNA-seq gene expression data of 40 samples. **d**) Hierarchical clustering based on RNA-seq gene expression data of 40 samples, with heatmap colors reflecting the pairwise correlation between samples. **e**) Aggregated ATAC-seq signals at *NANOG* locus in different cell types. **f**) Aggregated ATAC-seq signals at *GAD2* locus in different cell types.

1.3.2 Linking transcription factors with CREs

To establish the connection between *cis*- and *trans*-regulatory elements in our system, we performed TF footprint calling in the identified OCRs. By overlapping regions of local ATAC-seq signal depletion with TF motif instances, we found high-confidence TFBSs within the ATAC-seq OCRs of each cell type for all curated TF motifs in the 2018 JASPAR Core database, with the average length of a TFBS being 12 bp. The number of TFBSs identified for each TF motif in each cell type ranges from 10^2 to 10^4 (Fig. 1.3a), depending on factors such as the residence time of TF binding to DNA, the prevalence of given TF in the given cell type, and the coverage of OCRs in each cell type. In general, fewer TFBSs were detected iN-Glut and iN-DA than the rest cell types because of their relatively low coverage of OCRs. For CTCF, a TF with long DNA residence time

thanks to its multiple zinc finger domains that stabilize the interaction with the target DNA[47], our method detected $4\times10^3 - 2\times10^4$ footprints, which makes it rank ≥ 96% among all TFs across cell types. In contrast, the glucocorticoid receptor (GR, encoded by *NR3C1*) is known for its transient binding to DNA with binding kinetics two orders of magnitude faster than those of CTCF[47], [48]. Accordingly, the number of GR binding sites detected by our method makes it consistently rank at the bottom 2% quantile among all TFs across cell types.

Because of our footprint calling approach, the ATAC-seq cut count profiles averaged over all predicted TFBSs of a given TF generally exhibit characteristic signatures of TF footprints, with the binding sites protected from DNA cleavage while their immediate flanking regions show prominent cut signals (Figs. 1.3c-f). Using ENCODE ChIP-seq data measured in embryonic stem cells (H1-hESC) as reference, we further evaluated the performance of TFBS prediction within motif instances in iPSC for several TFs. The high ROC curves far away from the diagonal for all TFs considered (Fig. 1.3b) confirm that our adopted DNase2TF approach can sufficiently capture the TF binding events observed by ChIP-seq with low false positive rates despite the cell type difference between H1-hESC and iPSC. However, since some TF binding events can occur without leaving any footprints, the algorithm's choice of initial candidates as the local minima of the cut count profile led to incomplete ROC curves. In other words, DNase2TF cannot retrieve all ChIP-seq binding events in OCRs even at its lowest thresholding criteria. In fact, this phenomenon is commonly observed for other TFBS prediction algorithms, reflecting the limitation of footprint-based TFBS prediction methods[36].

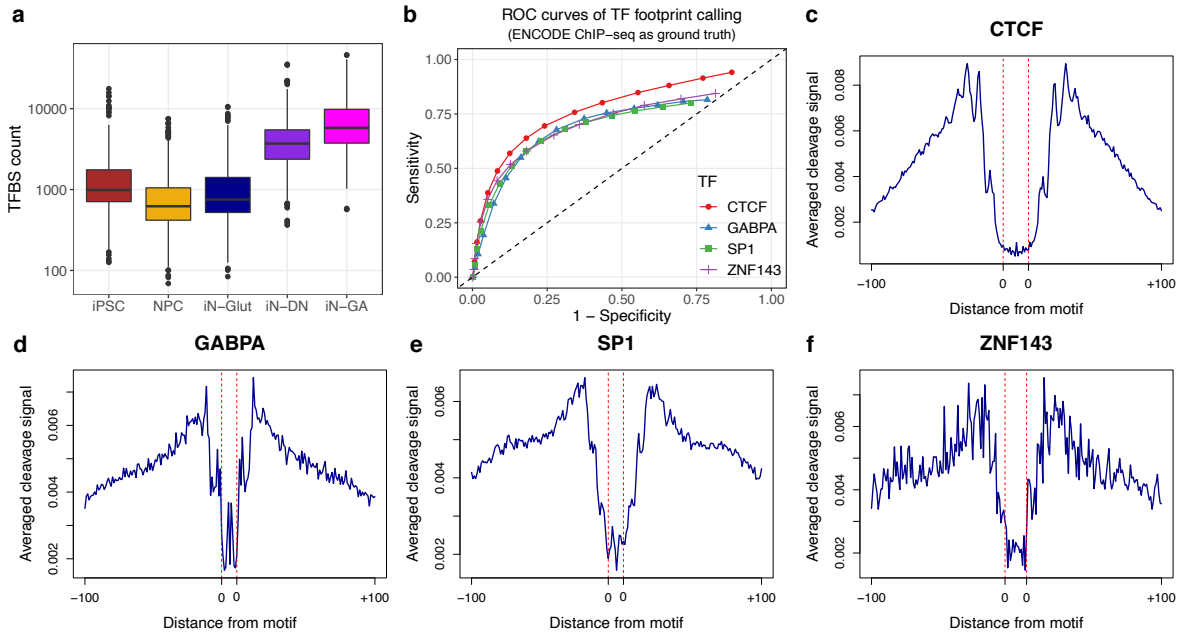

Figure 1.3 TF footprint calling results.

a) Distribution of the number of TFBSs for each of the 579 curated JASPAR TF motifs predicted in each cell type; the y axis is in log scale to display the wide range of TFBS counts. **b**) ROC curves of TFBS prediction in iPSC using ENCODE ChIP-seq data of matching TFs in H1- hESC as ground truth. **c**)-**f**) iPSC ATAC-seq cut count profiles at regions ±100bp of a predicted TFBS, averaged over all predicted TFBSs in the genome for c) CTCF (17672 sites), d) GABPA (2412 sites), e) SP1 (14416 sites), and f) ZNF143 (1138 sites). The region between two red dashed lines denotes the bound motif.

1.3.3 Linking CREs with target genes

Having identified the TF footprints located within CREs, our next step is to link CREs to target genes. Given the functional genomics data measured from this neurodevelopmental system, we decided to infer these functional links by statistically modeling the relationship between CRE activity and gene expression for target genes of interest.

We focused on 85 ASD risk genes with a score of 1 or 2 in the SFARI[42] and candidate ATAC-seq peaks within 100kb upstream or the intronic regions of these genes. We excluded any promoter peak from the following regression analyses as its regulatory target is clear, and the high correlation between its activity and target gene expression may mask the weaker effects of enhancer peaks that are less well-understood. Using elastic net penalized linear regression, we

identified from the candidate pool informative peaks that have the most effects on their target gene expression for each SFARI high-risk gene, utilizing ATAC-seq and RNA-seq data from all 40 samples. Starting from a candidate peak pool of ~44 (median) peaks per gene, we ended up with a selected peak pool of ~9 (median) peaks per gene, resulting in a selection rate ~7-fold (Fig. 1.4a). Despite being non-promoter peaks (Fig. 1.4c), the activities of selected peaks can explain their target gene expression response with an adjusted R^2 of ~0.8 (median), with the goodness of fit generally increasing with the number of selected peaks (Fig. 1.4b).

Further comparison between selected peaks and candidate peaks revealed that selected peaks tend to be close to the gene TSS instead of being intronic, and are enriched for H3K27ac marks and predicted TFBSs (Fig. 1.4e). We also examined the enrichment of specific types of TF footprints in these selected peaks under each cell type. Among the 5 cell types, 20 to 79 TFs were found to have significant footprint enrichment, including several TFs that are essential for neuronal development or differentiation, such as TFs of the *KLF* family[49], the *SOX* family[50], the *SP* family[51]–[53], *REST*[54], and *YY1*[55] (Fig. 1.4f, Table 1.1). Most of the enriched TFs are shared across multiple cell types. While iN-GA has the most types of TF footprint enrichment, 38 of which are unique to the iN-GA cell type, it is also the cell type with the greatest number of predicted TFBSs. Therefore, deeper investigation is needed to determine whether enrichment unique to cell types are due to the ATAC-seq coverage difference between cell types, or cell-type-specific regulatory activity.

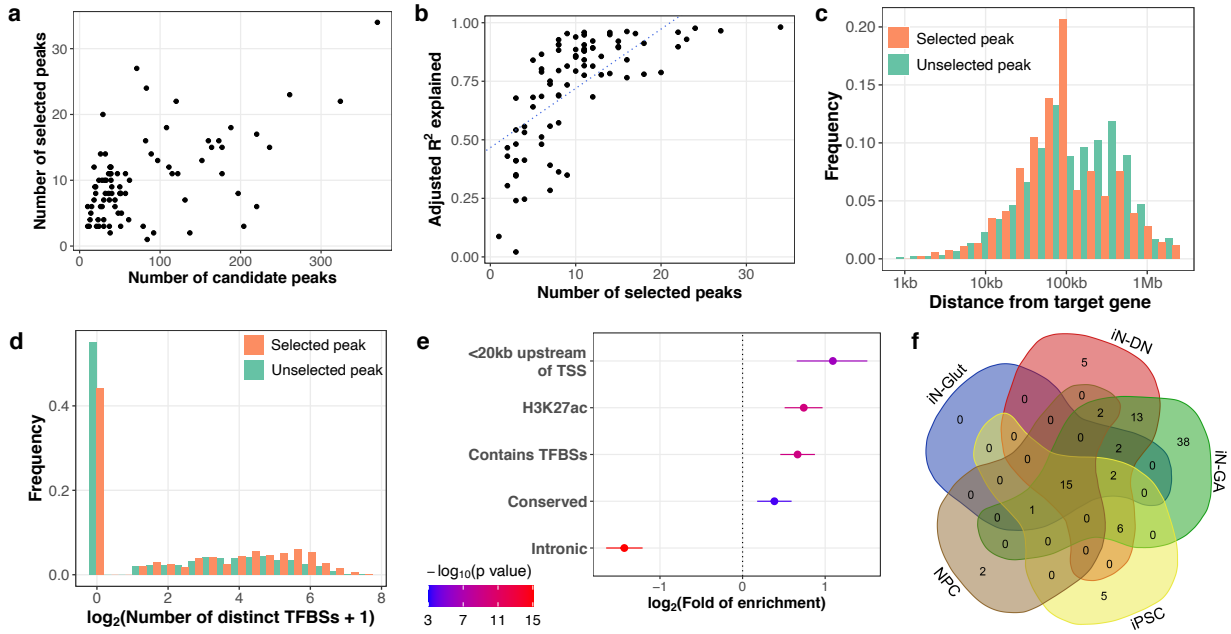

Figure 1.4: Elastic net selection of putative CREs (peaks) and characterization of selected peaks. **a**) The number of candidate CREs included in the model vs the number of CREs selected by elastic net regression per target gene. **b**) Relationship between number of selected CREs and their ability to explain target gene expression response as measured by adjusted R^2 in multiple regression per target gene. **c**) Distance of a peak from its potential target gene, for selected and unselected peaks in the candidate pool. *x*-axis is in log scale. **d**) The number of distinct types of TF motifs with predicted TFBSs in the peak in any cell type, for selected and unselected peaks in the candidate pool. *x*-axis is in log scale. **e**) The enrichment levels of different binary epigenetic features in selected CREs compared with the pool of candidate CREs; computed from simple logistic regression. **f**) Venn diagram reflecting the overlap of TFs that have predicted footprints enriched in elastic net selected peaks among cell types. TFBS enrichment was computed using the hyper-geometric test; significant TFs were obtained by thresholding FWER < 0.05 using the Bonferroni correction.

Cell types	Number of enriched TFs	TFs (motifs)
NPC, iN-DN, iN-GA, iN-Glut, iPSC	15	ZBTB7A, ELF1, ELK4, ETV4, FLI1, Gabpa, ERG, ELF4, ETS1, ELK1, NRF1, ELK3, ETV5, ETV1, FEV
iN-DN, iN-GA, iN-Glut, iPSC	2	EHF, ETV2
NPC, iN-GA, iN-Glut, iPSC	1	ETV3
iN-DN, iN-GA, iN-Glut	2	REST, YY1
iN-DN, iN-GA, iPSC	6	ETV6, Zfx, CTCFL, PLAG1, SP1, KLF16
NPC, iN-DN, iN-GA	2	TFAP2A, TFAP2B(var.2)
iN-DN, iN-GA	13	ZNF263, SP2, E2F6, NFATC2, MZF1, KLF5, CTCF, ELF5, SP3, KLF9, Rfx1, EWSR1-FLI1, ELF3
iN-DN	5	Arid3a, RORA(var.2), EGR1, NFAT5, Sox3
iN-GA	38	GLIS3, Tcf12, RBPJ, RARA::RXRA, TFAP2C(var.2), Pparg::Rxra, HNF4G, NR4A1, HINFP, PBX3, E2F4, ZEB1, ZIC1, Smad4, Myog, RELA, RORB, ASCL1, INSM1, Hes1, SP4, RREB1, TFAP2B(var.3), NR2F1, THAP1, TFAP2C(var.3), Klf1, Tcfl5, Klf12, TFAP2A(var.3), NR1A4::RXRA, TFDP1, ZIC3, NFIC::TLX1, Ascl2, RFX2, SP8, NFIC
iPSC	5	VDR, NA, RF, EGR2, IRF5
NPC	2	EN1, ALX3

Table 1.1: TFs (motifs) with footprints enriched in selected CREs in each cell type.

We inspected the selection of peaks by elastic net in detail for specific ASD risk genes. *ASH1L*, for example, has 50 candidate peaks within its introns or 100kb upstream region. Elastic net regression selected 8 peaks out of them, multiple regression using which can predict *ASH1L* expression level with an adjusted R^2 of 0.90. The activity of each selected peak also shows strong correlation with *ASH1L* expression level, with a marginal regression R^2 ranging from 0.37 to 0.87 (Fig. 1.5a), while unselected candidate peaks do not have marginal regression R^2 values above 0.48. Noticeably, given the high correlation we found among activities of candidate peaks for *ASH1L* (Fig. 1.5b) (potentially due to the consistent cell type difference across peaks), elastic net selected a parsimonious set of peaks that are also correlated with one another, as predictors in the final model. This reflects the "grouping" behavior of elastic net, where strongly correlated predictors tend to be included in or excluded from the model together[43]. Among the selected peaks, 5 of them contain predicted binding sites of TFs related to neural development (*e.g. SP8*[52],

NEUROD1, *SOX8*[50]) or associated with ASD risk (*e.g. MEF2C*[56]) across cell types (Fig. 1.5c), highlighting a complex regulatory pathway that connects *ASH1L* with key transcription regulators through enhancers.

For another ASD risk gene, *CHD8*, a chromatin remodeler, elastic net selected 8 putative regulatory peaks out of its 19 candidate peaks. Multiple regression using these selected peaks can explain *CHD8* expression level with an adjusted R^2 of 0.93. Again, elastic net selected groups of peaks that are highly correlated with each other (Fig. 1.6b). 6 of the selected peaks are distal, including 2 peaks at the promoters of other genes (Fig. 1.6c), which both have $R^2 > 0.8$ in marginal correlation with *CHD8* (Fig. 1.6a). One of them is positively correlated with the expression of both putative promoter genes: peak_385160 ~ *RAB2B* (R^2=0.83), peak_385160 ~ *TOX4* (R^2=0.75), suggests that peak_385160 may have regulatory control over multiple genes. Unexpectedly, the other peak (peak_385165) is anti-correlated with the expression of *METTL3* (R^2=-0.67), suggesting that it may not serve as a conventional promoter for *METTL3*. 7 of the 8 selected peaks have predicted TFBSs in at least one of the cell types, with some of the TFs in multiple peaks, such as those of the *SP* family (Fig. 1.6c). In contrast to *ASH1L*, the expression of *CHD8* is the highest in iPSC samples, and substantial iPSC TFBSs were detected in its neighboring peaks, which might be due to *CHD8*'s important role in the maintenance of pluripotency in embryonic stem cells[57].

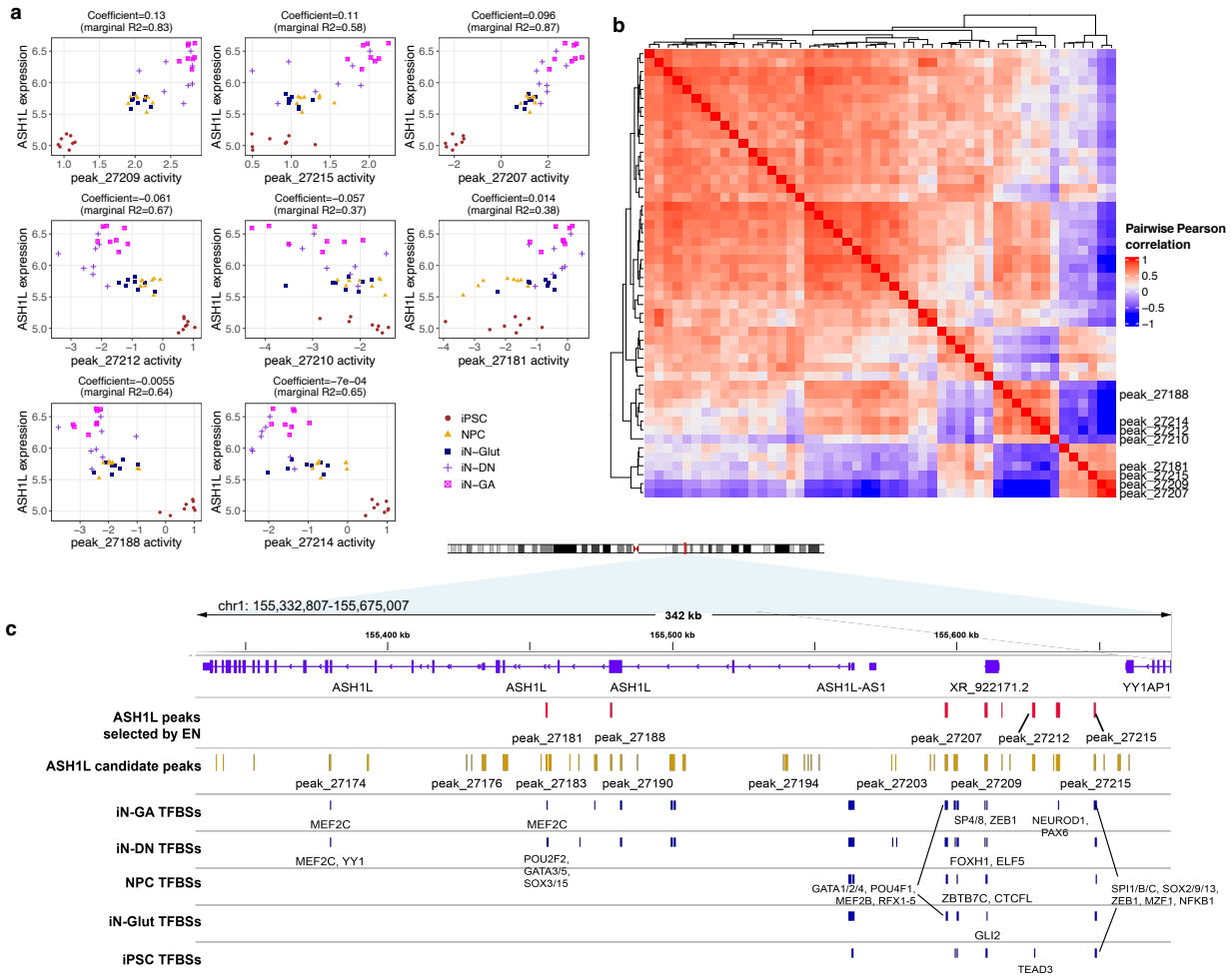

Figure 1.5: Elastic net selection of putative CREs that target *ASH1L*.

a) Correlation between activity levels of each selected CRE for *ASH1L* and the expression of *ASH1L* across 40 samples. Peaks are ranked by the absolute value of elastic net coefficient; R^2 from simple linear regression for each peak is also labeled. **b**) Pairwise correlation between the activity levels of 50 candidate peaks targeting *ASH1L*. **c**) Genomic track plot at the *ASH1L* locus, showing the locations of peaks selected by elastic net (red), candidate peaks for *ASH1L* (yellow), and predicted TFBSs in different cell types (last 5 tracks in blue). Each peak can contain multiple TFBSs, but due to limited space, only binding sites of selected TFs are labeled.

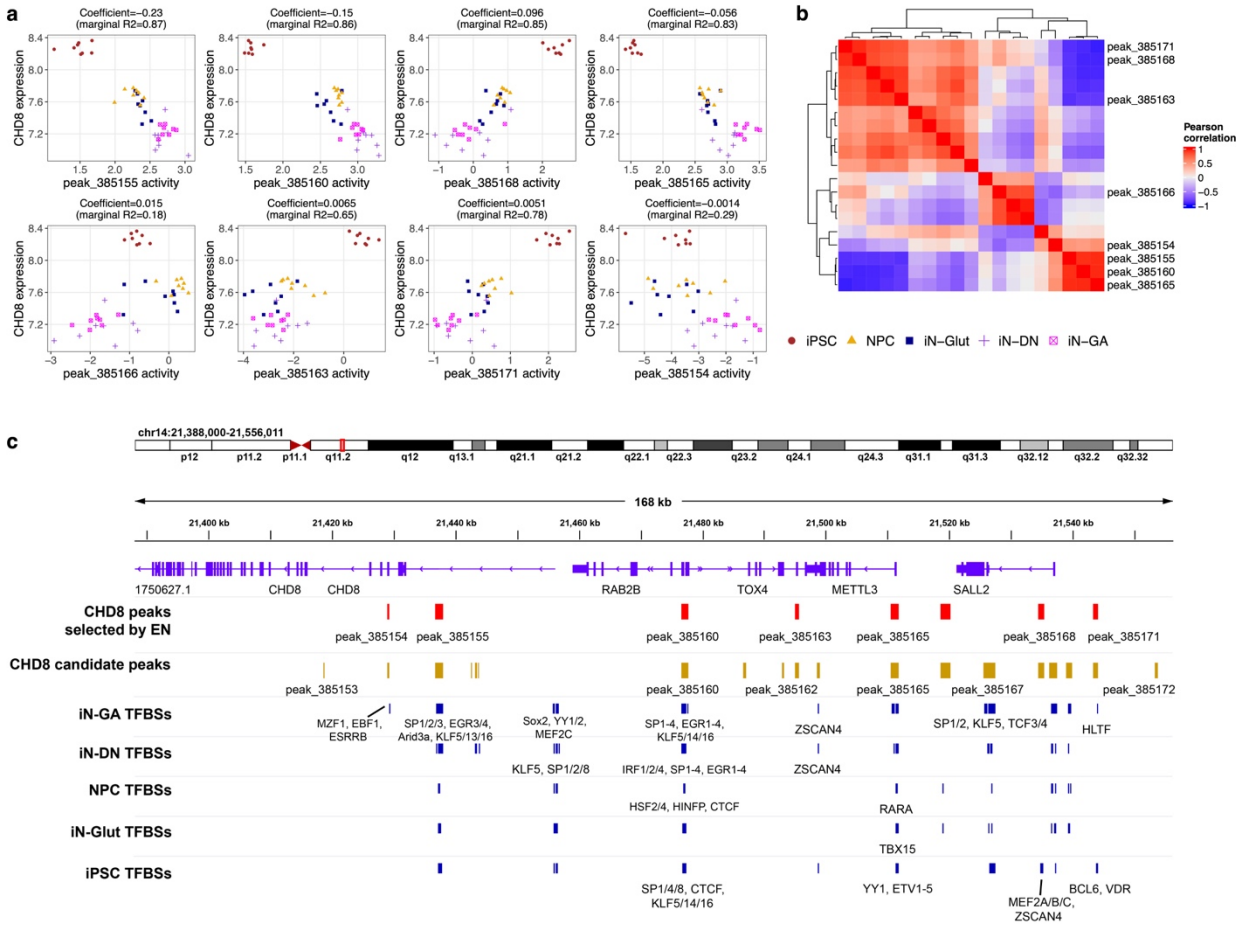

Figure 1.6: Elastic net selection of putative CREs that target *CHD8*.
a) Correlation between activity levels of each selected CRE for *CHD8* and the expression of *CHD8* across 40 samples. Peaks are ranked by the absolute value of elastic net coefficient; R^2 from simple linear regression for each peak is also labeled. **b**) Pairwise correlation between the activity levels of 19 candidate peaks targeting *CHD8*. **c**) Genomic track plot at the *CHD8* locus, showing the locations of peaks selected by elastic net (red), candidate peaks for *CHD8* (yellow), and predicted TFBSs in different cell types (last 5 tracks in blue). Each peak can contain multiple TFBSs, but due to limited space, only binding sites of selected TFs are labeled.

1.4 Discussion

In this preliminary work, we integrated the information from transcriptomics and functional genomics data, and used a series of computational approaches to establish connections from TFs to enhancers and to their target genes, mapping a preliminary gene regulatory networks at the transcriptional level in an iPSC-derived neuron model. This framework can be readily applied to other systems if matching RNA-seq and ATAC-seq data are available.

However, there remains room of improvement to increase the resolution and accuracy of resulting networks. Since the average length of peaks called from ATAC-seq is ~500bp, while the predicted TFBSs have an average size of ~12bp, a CRE typically contains multiple TFBSs. Complicated by the high degeneracy among TF motifs , *i.e.*, PWMs of TFs from the same family can be highly similar[58], a single CRE can contain up to 100 distinct types of TFBSs (Fig. 1.4d). Therefore, it is hard to identify the TFs that functionally bind to the CRE and where the important binding sites are located without additional information. Focusing on TFs with high expression levels may be one way to narrow down the list, as those are more likely to be functional in the system.

We used elastic net to link CREs to specific target genes. This approach, while effective in reducing the number of associated peaks, is generally not able to distinguish the functional CREs from a group of peaks highly correlated in activity. On the other hand, functional enhancers are typically associated with epigenetic characteristics such as H3K27ac marks, physical interaction with promoter, and bound by TFs. Enrichment for some of these features is indeed observed in peaks selected by elastic net regression. Therefore, we can leverage the functional information in sequences to facilitate the selection of CREs, prioritizing those with matching epigenetic features over others. Bayesian variable selection regression[59] is an attractive solution for this purpose, as the epigenetic information can be incorporated into the model as sparse priors imposed on effect sizes of predictors. Similar ideas have been applied to prioritize functional disease variants from GWAS, as seen in the category of statistical fine-mapping approaches[60].

It is also important to keep in mind that association does not imply causation, and experimental studies are needed to confirm the validity of these computationally inferred regulatory networks. Enhancer activity can be validated through luciferase assay *in vitro*, or more

recently, via high-throughput techniques such as Massively Parallel Reporter Assay (MPRA)[61]. Promoter capture Hi-C[12] can identify enhancers that directly interact with the promoter of a gene over long ranges. The regulatory effects of CREs on gene expression or cellular phenotypes can be directly assessed *in vivo* through CRISPR/Cas9 genome editing[16]. Only with the support of orthogonal experimental validation can one draw a functional link between regulatory elements and their target genes with certainty.

CHAPTER 2: STATISTICAL ANALYSIS OF CHROMATIN ACCESSIBILITY VARIANTS TO ELUCIDATE REGULATORY GENETICS OF SCHIZOPHRENIA[1]

2.1 Introduction

In the previous chapter, we have observed how ATAC-seq paired with RNA-seq, can help us identify potential *cis*-regulatory elements (CREs) and the disease-risk genes they are associated with. However, due to the lack of computational or experimental screening, the predicted functions of these CREs were not validated. In addition, as the prediction of CREs does not inform the effects at the genetic variant level, the causal variants underlying diseases remain unknown. Genome-wide association studies (GWAS) have mapped a large number of genetic variants associated with neuropsychiatric disorders, most of which reside in noncoding regions of the genome[14]. Still, it remains challenging to pinpoint the causal variant among these GWAS risk variants due to the prevalent linkage disequilibrium (LD) patterns in the human genome, as SNPs in high LD with the causal SNP can be statistically associated with the trait, generating spurious signals or even masking the signal of the causal SNP[60].

In this work, we established a novel framework to address these challenges, combining the rich regulatory information in neuronal ATAC-seq data with the disease relevance of genetic variants summarized in GWAS of neuropsychiatric disorders. In contrast to most existing studies performed on adult brain samples[62], we focused on a neurodevelopmental system consisting of major subtypes of neuronal cells derived from human induced pluripotent stem cells (iPSCs),

[1] Much of this chapter contains material from the paper: Zhang *et al.*, "Allele-specific open chromatin in human iPSC neurons elucidates functional disease variants," *Science*, vol. 369, no. 6503, pp. 561–565, 2020, doi.org/10.1126/science.aay3983.

which may provide functional information during early neurodevelopment. As chromatin accessibility strongly influences gene expression during neurodevelopment, we set out to investigate the extent to which genetic variants alter chromatin accessibility in these neuronal cell types. Based on ATAC-seq open chromatin profiling, we mapped thousands of genetic variants exhibiting allele-specific open-chromatin (ASoC) in each cell type. We evaluated the cell-type specificity of these ASoCs, and validated the functional effects of them via their enrichment for brain-specific quantitative trait loci (QTLs) and their alteration of transcription factor (TF) binding. The disease relevance of these ASoC variants was then assessed using GWAS summary statistics of brain-related disorders and traits, and top 20 putative causal variants for schizophrenia (SZ) were prioritized via a Bayesian fine-mapping strategy that incorporated multiple important genomic features, such as neuronal ASoCs, as functional priors.

Given these highlighted variants, functional validations are needed to confirm whether they and their underlying sequences are responsible for the downstream phenotypes such as gene expression and cellular functions, which ultimately affect the disease trait. One way to establish this causal link is to directly introduce genetic perturbation at the sequence of interest, and evaluate the phenotypic outcomes of the perturbation in cells, which is enabled by the CRISPR/Cas9 genome editing system. Through CRISPR, a regulatory sequence can either be entirely disrupted by the Cas9 nuclease, or repressed in its regulatory activity by a nuclease-deactivated Cas9 (dCas9) in CRIPSRi[16].

In the second half of the work, we centered on the top 20 ASoC SZ-risk SNPs and set to validate their regulatory functions using CRISPR genome editing. We first used CROP-seq (CRISPRi followed by droplet-based single-cell RNA-seq)[22] to target the tagged ASoC sequences in a pooled experiment. Differential gene expression analysis was then carried out to

screen for genes that are *cis*-regulated by these SZ-risk regulatory sites. Next, independent CRISPRi followed by qPCR assay was used to validate the *cis*-genes identified in the CROP-seq screen. Finally, CRISPR editing at the SNP level confirmed the allelic effects of two variants on the local regulatory activity and on the expression level of the corresponding *cis*-target genes.

Overall, our work provided a snapshot of the neuronal ASoC landscape and demonstrated that ASoC data in iPSC-derived neurons provide an effective means to interpret functional effects of variants in the neurodevelopment context, and to aid genetic discovery of neuropsychiatric traits. While the discovery framework we established is a combination of state-of-art experimental and computational techniques, the focus of this chapter is on the computational side.

2.2 Methods

2.2.1 Mapping of allele-specific open chromatin (ASoC) variants

We first set out to identify functional variants affecting chromatin accessibility, which we hypothesize are more likely to influence gene expression. We focused on ASoC variants displaying allelic imbalance in ATAC-seq reads at heterozygous SNP sites in a neurodevelopmental context. To this end, we expanded the study system used in Chapter 1 to include more informative samples. 20 reprogrammed iPSC lines were obtained from 20 subjects (Table 2.1) in the Molecular Genetics of Schizophrenia (MGS) cohort[63], including the 8 subjects used in Chapter 1. These individuals were selected for this study because they are "super-heterozygous" for SZ risk SNPs. They are enriched for heterozygous SZ GWAS index SNPs at 70 SZ loci (out of 108 in total[55]), and provide sufficient power to detect ASoC with a minimum ATAC-seq read depth of 20 at these 70 loci[64]. As mentioned in Chapter 1, these iPSC lines were differentiated into neural progenitor cells (NPC), early-stage (day-15) glutamatergic (iN-Glut), GABAergic (iN-GA), and dopaminergic (iN-DN) neurons. But in addition to the original 8 lines ("core 8") per cell type,

ATAC-seq and RNA-seq were performed on 12 additional lines in NPCs and iN-Glut individually, resulting in 20 sequenced samples for each of these two neuronal cell types.

Cell ID	iPSC ID	Sex	Age	Condition
01C08162	CD0000002	F	48	case
04C27190	CD0000003	M	46	case
04C28905	CD0000004	M	19	control
04C37433	CD0000005	M	65	control
05C38571	CD0000006	F	55	control
05C39664	CD0000007	M	59	control
05C43356	CD0000008	M	80	control
05C43758	CD0000009	F	32	control
05C45915	CD0000010	M	52	case
05C46807	CD0000011	F	58	case
05C46837	CD0000012	M	42	case
05C48054	CD0000013	M	86	case
05C49221	CD0000014	F	33	control
06C52191	CD0000015	M	71	control
06C52565	CD0000016	M	57	case
06C52573	CD0000017	F	45	case
06C53368	CD0000018	F	42	case
06C54426	CD0000019	M	41	control
07C71166	CD0000020	M	45	case
07C65853	CD0000021	M	29	case

Table 2.1: MGS subjects used for generating iPSC lines

ASoC variant calling

Given the ATAC-seq reads, SNP variants were called using GATK[65] (version 4.0), and then tested for allelic imbalance of chromatin accessibility using a binominal test (Fig. 2.1). Briefly, WASP[66] was used to account for the mapping bias to reference alleles in aligned reads before variants were called using the discovery mode of HaplotypeCaller. Each sample was processed individually, and only heterozygous SNP sites with corresponding rs# records found in dbSNP v150 were retained. To maximize the power of ASoC detection, for each called SNP site, we pooled reads from samples within the same cell type that are heterozygous at the site. This sample

pooling strategy is justified by the high concordance of allele-specific effects within a cell type across individuals[64]. Finally, the VCF files were generated and filtered such that only biallelic SNP sites (GT: 0/1) with a read depth (DP) $\geqslant$ 20 and minimum reference or alternative allele count $\geqslant$ 2 were retained. A two-sided binomial test was then carried out for each SNP site, assuming a null model where the number of trials is DP and the probability of success is 0.5. Multiple testing correction was performed using the Benjamini-Hochberg procedure for SNPs that were tested in a cell type, and ASoC SNPs for this cell type were obtained under an FDR cutoff of 0.05.

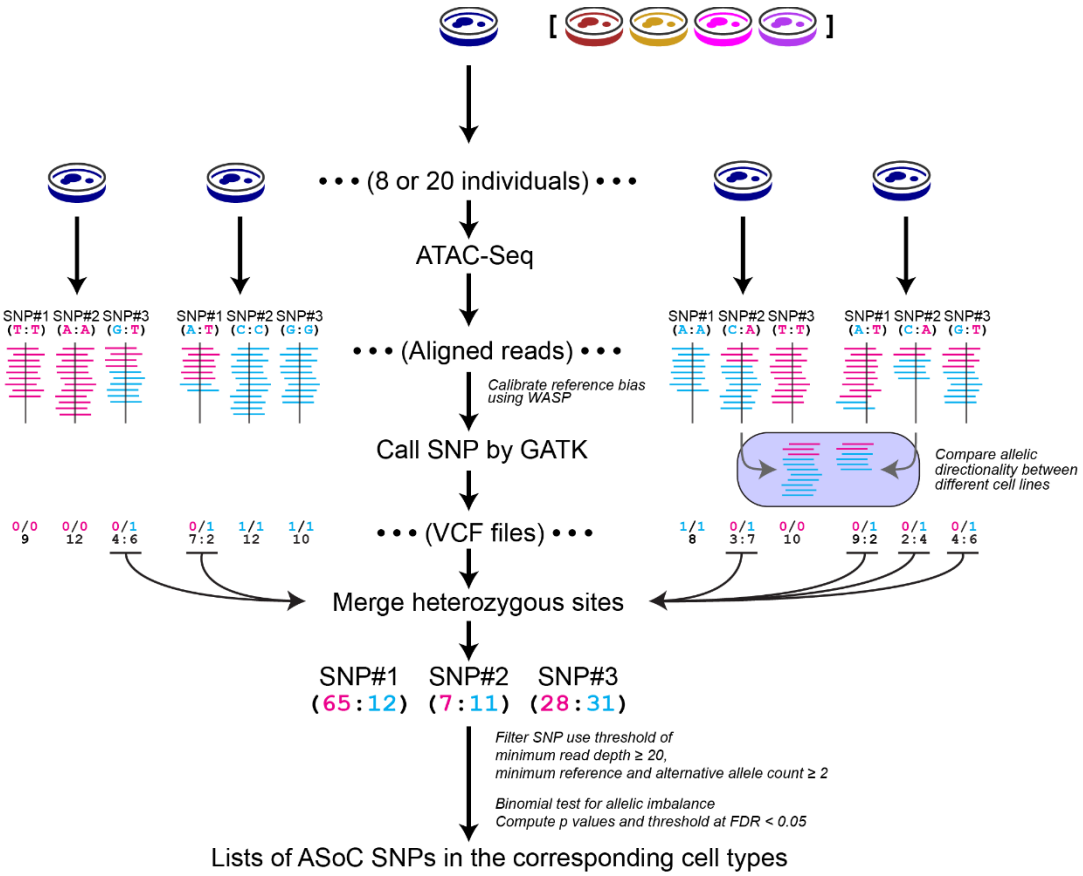

Figure 2.1: Workflow of ASoC SNP calling in ATAC-seq reads of a given cell type.

Cell-type specificity and sharing of ASoC SNPs

We next investigated the sharing patterns of ASoC variants among cell types. In general, we defined an ASoC SNP to be cell-type-specific if it has an FDR < 0.05 from the imbalance test in its own category, while its nominal p-value > 0.05 in all other cell types. We defined cell-type-shared SNPs as those with FDR values < 0.05 in all shared groups.

We used Storey's π1 analysis[67] to further evaluate the pairwise sharing of ASoC SNPs among cell types. For each pair of 'leading' and 'matched' cell types, we obtained the ASoC SNPs (FDR < 0.05) from the leading cell type, and estimated the proportion of non-null tests (π1) from the distribution of nominal p-values of these SNPs in the matched cell type if the test statistics exist. Note that if an SNP has a DP < 20 or was not called in the matched cell type, it was automatically counted toward the null proportion. Therefore, the π1 proportion was further scaled to reflect the proportion of ASoC SNPs in the leading cell type that are shared by the matched cell type.

2.2.2 Characterization of ASoC variants

We used statistical analyses to characterize the functions of these called ASoC variants in several aspects, including their enrichment in functional regions of the genome, their relationships with transcription factor (TF) binding sites, and their relevance to neuropsychiatric disorders.

Enrichment of ASoC SNPs in annotated genomic regions

The definitions of chromatin state were assembled using an imputed 25-state model derived from individual #E081 of fetal brain tissue by the Roadmap Epigenomics Project[68]. We categorized chromatin states 1–4 as "promoters" and chromatin states 9–19 as "enhancers".

For sets of ASoC SNPs that are cell-type-specific or shared by three cell types, we analyzed their enrichment within the above epigenetically annotated regions using a one-sided binomial test

Binom(x; n, p), where x is number of SNPs that fall into the designated epigenetically annotated features, n is the total number of SNPs under consideration, and p is the ratio between the total length of the designated epigenetically annotated features and the size of human genome.

Enrichment of ASoC SNPs in TF binding sites

To gain insight into the regulatory mechanism of ASoC, we then examined whether ASoC SNPs in each cell type were enriched at specific TF-binding sites (TFBSs). TF-binding footprints were called within each cell type using ATAC-seq data following the approach described in Chapter 1 but for 522 Homo sapiens TF motifs in the JASPAR database (2018 release), generating two sets of TFBS for iN-Glut-20 and NPC-20 samples, in addition to the five sets of TFBSs predicted from "core-8" samples per cell type. The enrichment of ASoC in TFBSs in each cell type can then be assessed as follows. Using the ~1M TFBSs predicted in 20 iN-Glut cell lines, for example, we evaluated the enrichment level of iN-Glut-20 ASoC SNPs (FDR < 0.05) vs iN-Glut-20 non-ASoC SNPs (FDR > 0.05) that are located in the predicted TFBSs of a given motif using Fisher's exact test.

Correlation between allelic imbalance of chromatin accessibility and TF motif disruption

We next utilized the observed chromatin accessibility levels at these ASoC alleles to test if a motif/TF has a role in driving chromatin accessibility. For such a TF, we would expect that genetic perturbation of its binding site would lead to a matched change of chromatin accessibility. For instance, an allele disrupting the motif of a pioneer factor is expected to reduce chromatin accessibility. This analysis is analogous to a Mendelian randomization[69] approach where genetic variants are used to test the causal effect of an exposure (in our case, TF/motif) on the outcome (chromatin accessibility).

Specifically, for a given cell type, we first obtained all the ASoC SNPs that are located inside the identified TFBSs of any of the 522 JASPAR TF motifs. Next, we use the function motifbreakR() from R package motifbreakR[70] to evaluate the level of disruption the alternative allele of each SNP has on any possible motif that matched its surrounding sequence with a p-value filtering threshold of 2.5×10^{-4}. The effect of disruption can be either positive or negative, corresponding to increased or decreased TF binding affinity compared with the sequence with reference allele, respectively. Only SNPs with 'strong' effects on their underlying motifs according to motifbreakR were kept for further analysis, and their motif disruption scores were quantified as the difference between 'scoreAlt' and 'scoreRef' from motifbreakR() results. Then, for each given TF motif and all ASoC SNPs that 'strongly' disrupted its TF-binding motifs, we fit a linear regression model with an offset of zero between the motif disruption scores of these SNPs and their allelic imbalance levels. The latter was quantified by taking the log2 of the ratio between alternative and reference ATAC-seq reads at each of the SNP loci. Polarized -log10 of regression p-values were used to indicating both the significance and direction of the association between motif disruption and allelic imbalance of chromatin accessibility.

Enrichment analyses of ASoC SNPs in brain molecular QTLs

To test whether brain-associated SNPs from quantitative trait loci (QTL) are enriched in the functional genomic annotations we generated, we performed SNP-based enrichment analysis using a Bayesian hierarchical model (TORUS)[71]. TORUS incorporates the enrichment of functional annotations through a logistic link on the spike-and-slab prior on QTL effect sizes (β_j's):

$$\beta_j \sim \left(1-\pi_j\right)\delta_0(\cdot) + \pi_j g(\cdot),$$

$$\log\frac{\pi_j}{1-\pi_j} = \alpha_0 + \sum_k \alpha_k d_{jk}.$$

Here π_j is the prior inclusion probability for the j-th SNP in a certain locus, d_{jk} is the value of the k-th annotation for SNP j, and α_k is the enrichment parameter, or the log odds ratio, of the k-th annotation. Here, we investigated the enrichment of brain-specific QTL data of expression, methylation and histone acetylation[72] each in one functional annotation: iN-Glut-20 ASoC SNPs, assigning a binary status to each SNP, with 1 being an ASoC SNP and 0 otherwise.

2.2.3 Differential expression analysis for single-cell RNA-seq with CRISPR screen

Given the strong enrichment of ASoC SNPs for SZ-risk variants, we selected the top 20 non-MHC (major histocompatibility complex) ASoC SNPs that are also GWAS SZ index SNPs or their LD proxies ($r^2 \geqslant 0.8$) (Table 2.2) to assess the regulatory potential of their underlying sequences and the likely *cis*-target genes. We adopted a modified CROP-seq[22] approach to repress the activities at these sites of interest in cells and screen for their effects on gene expression (Fig. 2.2). Briefly, for each SNP, three gRNAs were designed to target sequences ±150 bp of the SNP site on both the forward and reverse strands. Three NPC lines stably expressing dCas9/KRAB were transduced with a lentiviral gRNA library containing all these designed gRNAs under a low multiplicity of infection (MOI = 0.2) setting to maximize the number of cells infected with a single gRNA. Transduced cells were processed with the 10x Genomics Chromium single-cell 3' mRNA platform and sequenced, generating approximately 450 M reads in total.

10x Genomics Cellranger v2.1.2 was used to process raw sequencing data, where sequenced reads were aligned to the human GRCh38/hg38 genome with spike-in gRNA sequences as artificial chromosomes (20 bp gRNA sequence and 250 bp downstream plasmid backbone per each gRNA) using STAR v2.5.1. A digital gene expression matrix was then constructed based on per-gene UMI count. Cells with more than 500 genes and 11,000 UMIs detected were retained (4,099 out of 4,144 cells) for subsequent analysis, resulting in a median of 28.5k total UMIs and a

median of 5k genes detected per cell. We assigned a cell with a unique RNA when the UMI count of its dominant gRNA is at least three times more than the sum of UMI counts from all other gRNAs. Under this criterion, 2,522 cells (60.8%) were uniquely assigned with one type of gRNA, and were retained for further analysis.

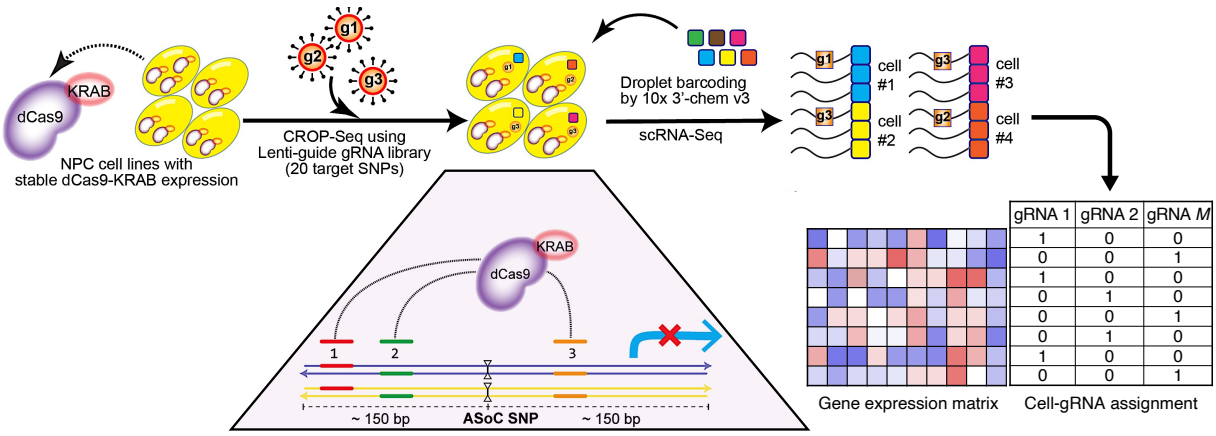

Figure 2.2: The design of a modified CROP-seq approach for multiplexed CRISPR/dCas9 epigenomic perturbation at ASoC SNP sites with scRNA-seq readout

After quality control and preprocessing, we performed differential expression (DE) analysis on the single cell gene count data using the edgeR package, where a generalized linear model was fit to the data and DE genes were detected using a quasi-likelihood F-test test (edgeR-QLF)[73]. This test was selected because it has been shown to perform better than most other methods in identifying DE genes in scRNA-seq experiments with small cell numbers (range of 6-400 cells)[74]. While each sequence was targeted by 3 gRNAs, we found that the efficiencies of each gRNA often differed from the other (as measured by the change in expression of their targeting genes), so we analyzed the cells designated to each gRNA individually. Cells containing the negative control gRNAs (3 targeting EGFP and 2 scrambled sequences) were used as the control group in DE analysis. Genes were included if they have CPM > 30 in more than 20% of cells.

When DE analysis is applied to scRNA-seq data, however, it is commonly observed that the distribution of test p-values across genes have inflation close to 0, resulting in an excessive number

of false positives above the imposed level[74]. To correct for the p-value inflation observed in our DE test results, we performed permutation tests to obtain empirical p-values that are calibrated. Specifically, for each gRNA condition, we permuted the labels of cells randomly so that they no longer correspond to their original treatment conditions and performed the same edgeR DE test for each gene. 30 rounds of random permutations were carried out like this, and the p values resulted from tests of all genes over all permutations were pooled together to form an empirical null distribution. We then used this null distribution to assign an empirical p-value to each gene, computed as the proportion of p values in the null distribution that are less than or equal to the observed p-value from the original DE test. Given the relatively small numbers of cells and limited power, we only considered *cis*-genes within 500 kb of each targeted SNP and used an empirical p-value threshold of 0.05 to identify differentially expressed *cis*-genes.

After the DE *cis*-genes were obtained, we further assessed the enrichment of genes that were tested as transcriptionally repressed among all discovered *cis*-genes using Fisher's exact test. For this part, we defined our *cis*-gene discovery list by varying the cutoff on gene distance from target SNPs from 50 kb to 1Mb, and conducted the enrichment test under each cutoff.

2.3 Results

2.3.1 OCR and ASoC landscapes in iPSC-derived neuronal cell types

Because open chromatin often overlaps with regulatory DNA sequence[75], [76], localization of disease risk variants within open chromatin regions (OCRs) can help prioritize putative functional noncoding risk variants for neuropsychiatric disorders[40], [77]–[79]. However, not all the variants within OCRs are functional, and as a result, the enrichment of GWAS signals in OCRs is often modest[78]. It thus remains a challenge to precisely identify functional risk variants for neuropsychiatric disorders.

We first set out to identify functional variants affecting chromatin accessibility, which we hypothesize are more likely to influence gene expression in a neurodevelopmental context. We mapped the chromatin accessibility profiles in major iPSC-derived neuronal subtypes using ATAC-seq, and then conducted a comprehensive ASoC mapping in each cell type to seek a direct functional readout of noncoding risk variants for neuropsychiatric disorders.

To pick informative samples for our ASoC discovery, we derived iPSC lines from 20 subjects (Table 2.1) from the MGS cohort, chosen for their enrichment for heterozygous GWAS index SNPs (~70/108 SZ loci). These iPSC lines were differentiated into 4 neuronal cell types, and ATAC-seq and RNA-seq were performed within each cell line sample. Principal component analysis (PCA) of both ATAC-seq and RNA-seq samples showed cell-type-specific clustering (Fig. 2.3a,b). As expected, OCRs of iPSC-derived neural cells were more similar to those of fetal brains[80] and cortical organoids[81], than to PsychENCODE adult brains[78] (Fig. 2.3a). While neuronal OCRs overlapped with 45-55% of the PsychENCODE peaks (n = 117,935), they only accounted for ~20% of our neuronal OCRs, consistent with the difference in enhancers observed between fetal and adult cerebral cortices[81].

We then called ASoC variants in each cell type from the ATAC-seq reads (Fig. 2.1). To increase power, we adopted a sample-pooling approach[82], [83] after confirming the inter-individual concordant directionality of the allelic imbalance of candidate ASoC SNPs within each cell type. Within each set of core-8 cell lines, we identified 920-2,392 ASoC SNPs (FDR < 0.05) in each cell type, with most OCR peaks containing a single ASoC SNP (Fig. 2.3c).

Comparison across 5 cell types revealed abundant cell-type-specific ASoC SNPs (Fig. 2.3d). Using Storey's $\pi 1$ analysis[67], we estimated pairwise ASoC SNP sharing and found neuronal cell types shared a low percentage with iPSC (10-20%; Fig. 2.3e). For shared ASoC SNPs across cell

types, the direction of allelic imbalance was well correlated (R = 0.77-0.95). However, even within neuronal cell types, ASoC sometimes differed substantially (30-70% sharing) (Fig. 2.3e), suggesting high cell-type specificity of ASoC.

We inspected in depth whether the observed cell-type specificity of ASoC is driven by cell-type-specific OCRs or different SNP effect sizes across cell types. About half of the neuronal ASoC SNPs do not have enough ATAC-seq coverage in iPSC (Fig. 2.3f). For neuron-specific ASoC SNPs that have sufficient read depth in iPSC, we found that most of them have insignificant allele imbalance in iPSCs despite being in strong OCRs (Fig. 2.3g), reflecting cell-type-specific allelic effects. In fact, about 50% of OCRs containing neuron-specific ASoC SNPs are more active in their corresponding neuronal cell type, while the other half showed comparable or even higher accessibility in iPSCs (Fig. 2.3h). These results suggest that variations in both OCR intensities and SNP effect sizes across cell types contribute to driving cell-type-specific ASoC, highlighting cell-type-specific regulation in the absence of chromatin state changes.

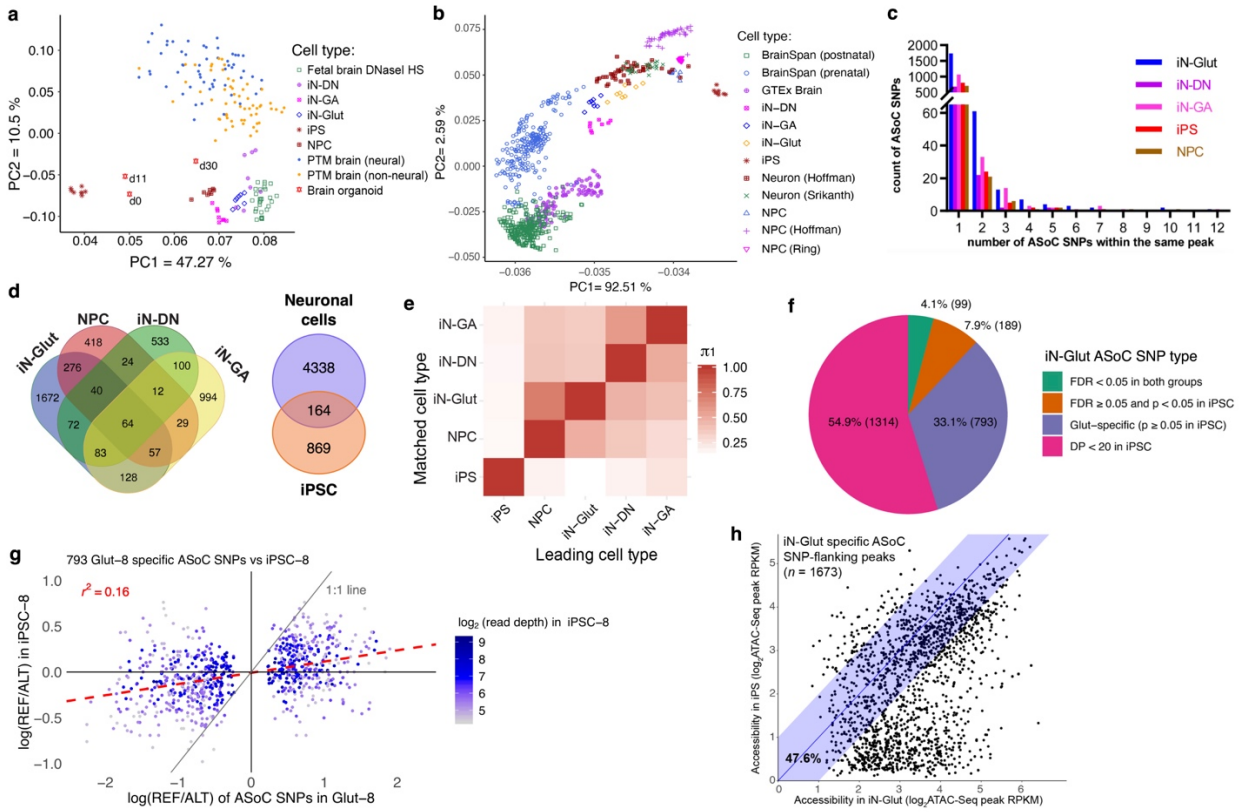

Figure 2.3: Mapping of ASoC variants in iPSC-derived neuronal cell types.
Data shown are all from core-8 samples. **a**) PCA of OCR intensities shows a higher similarity between iNs and fetal brains/day-30 organoids than postmortem (PTM) adult brain. All were ATAC-seq samples except for the fetal brain. **b**) PCA analysis of RNA-seq samples of the core-8 cell lines in comparison with multiple publicly available RNA-seq datasets on iPSC-derived neural cells or brain tissues (see[64] for details). **c**) Distribution of ASoC SNPs within peaks in each cell type. **d**) Venn diagrams of ASoC SNP overlap in different cell types. **e**) Pairwise $\pi 1$ estimation of ASoC sharing across cell types. ASoC SNPs were ascertained in the leading cell type, and $\pi 1$ was estimated in the matched ones. **f**) Sub-categories of ASoC SNPs in iN-Glut broken down by their accessibility in iPSC. **g**) Correlation between allelic ratios in iN-Glut and in iPSC for iN-Glut ASoC SNPs that have a DP $\geq$ 20 in iPSC. Data points are colored based on their DP in iPSC; the dashed red line represents a linear model fit using all data points. **h**) Chromatin accessibility in iN-Glut vs. iPSC for OCR peaks flanking iN-Glut-specific ASoC SNPs (DP < 20 or p-value > 0.05 in iPSC). Peaks that fall into the shaded area have comparable (within 2-fold difference) accessibility between cell types.

2.3.2 ASoC SNPs are enriched for functional characteristics

We next examined the genomic/epigenomic features of these ASoC SNPs. As expected, ASoC SNPs were enriched in brain promoters and enhancers (Fig. 2.4a). However, compared with shared

ASoC SNPs, cell-type-specific ones showed about 2-fold higher enrichment in enhancers (Fig. 2.4a).

As the number of detected ASoC SNPs linearly increases with sample size[64], we used ASoC SNPs discovered in cell types with 20 cell lines (NPC-20 and iN-Glut-20) to maximize our power in our downstream analyses. In total, we identified 5,611 ASoC variants in iN-Glut-20 and 3,547 ASoC variants in NPC-20, with 1,690 shared between them.

To validate the functional relevance of ASoC to brain gene regulation, we first jointly analyzed our data with other studies of genetic variants associated with brain-specific quantitative trait loci for gene expression (eQTL), histone modification (haQTL), and DNA methylation (meQTL)[72]. We found that iN-Glut ASoC SNPs were enriched (> 30-fold) for putatively causal variants underlying brain eQTL, haQTL, and meQTL (Fig. 2.4b). These results thus support the role of ASoCs in regulating brain gene expression.

To gain insight into the regulatory mechanism of ASoC, we mapped the TF-binding footprints from ATAC-seq. In iN-Glut-20 neurons, out of the 5,611 ASoC SNPs, 1,802 (32%) were found inside TFBSs, representing a 1.6-fold enrichment (vs. non-ASoC SNPs, Fisher's exact test, p-value = 2.6×10^{-58}), and most ASoC SNPs were within 200 bp of their nearest footprints (Figs. 2.4c,d). Footprint analysis using the core-8 samples in each cell type gave similar results[64]. To test whether TF-motif disruptions by ASoC SNPs near TF-binding footprints caused matched changes of chromatin accessibility, we performed correlation analysis between imbalance of accessibility and disruption of TF motif binding at these loci. In iN-Glut-20 and NPC-20 cells, we found 48 TFs with SNP motif-disruption scores either positively (*e.g.*, *SOX9*, a known pioneer TF) or negatively correlated with the allelic imbalance of ASoC (Figs. 2.4e,f). These results suggest that altering TF-binding is an important mechanism for the allelic effect of ASoC variants.

We further examined whether ASoC SNPs in each cell type were enriched at TF-binding sites of specific TFs. We found somewhat cell-type-specific patterns of enriched TFBSs, and identified the enrichment of some crucial TFs including *POU3F2/3*, *EGR3/4*, *NHLH1*, and *SP4* in iN-Glut (Fig. 2.4g), with *SP4* being a SZ risk gene[55]. These results suggest that ASoC SNPs may affect the cell-type-specific binding of TFs important for cell fate commitment and neurodevelopment.

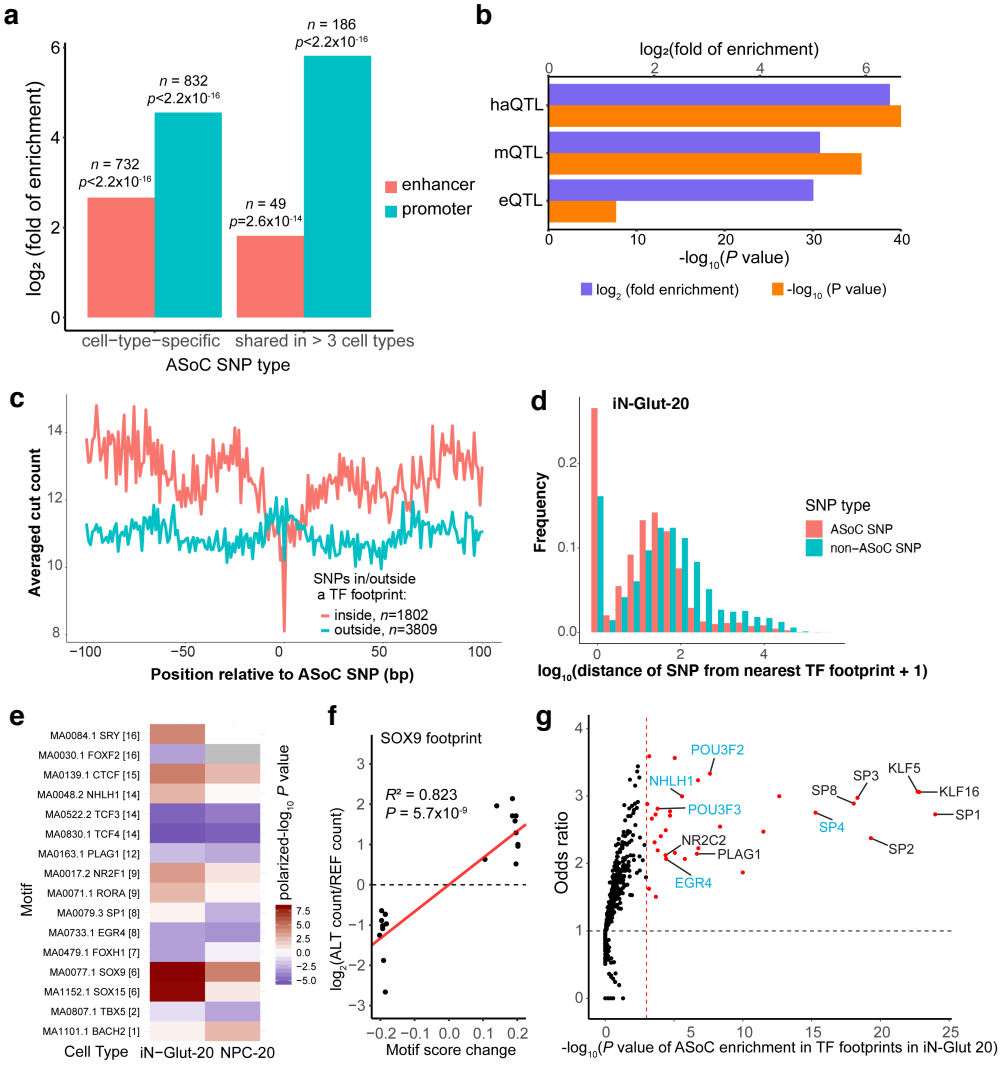

Figure 2.4: Characteristics of ASoC SNPs.

a) Enrichment of ASoC SNPs in the chromatin-state- based annotations of promoter and enhancer. **b**) Enrichment of iN-Glut-20 ASoC SNPs for brain QTLs. **c**) Averaged ATAC-seq cleavage profiles around ASoC SNPs inside (red, n = 1,802) or outside (blue, n = 3,809) predicted TF footprints. The ATAC-seq cleavage profiles were generated by piling up the 5' ends of ATAC-seq reads in 200 bp windows centered around the ASoC SNPs. **d**) Distribution of SNP distance from their nearest TF footprint for ASoC SNPs and non-ASoC SNPs in iN-Glut-20. The distance (bp) is defined as from the SNP to the nearest end of the

(Figure 2.4 continued)
corresponding footprint and presented in in log scale; the average size of a predicted footprint is 12 bp. **e**) Selected TFs with significant correlation (FDR < 0.05) between motif disruption score and ASoC allelic imbalance. **f**) *SOX9* as an example of putative chromatin activator showing positive correlation between motif disruption score and ASoC allelic imbalance. **g**) Enrichment of neural ASoC SNPs in footprints of specific TFs. Blue: TFs important for neurodevelopment.

2.3.2 Validations of ASoC SNP functions with CRISPR screens

We further assessed the utility of ASoC for inferring functional noncoding GWAS risk variants for neuropsychiatric disorders. Among the 5,611 ASoC SNPs in iN-Glut-20, 21 were found to be SZ-associated SNPs or their linkage disequilibrium (LD) proxies at 17 independent SZ loci, reflecting a 5.2-fold of enrichment compared with all heterozygous SNPs in these samples (Fisher's exact test, Fig. 2.5a). Using TORUS[71], a statistical method that accounts for the uncertainty of GWAS risk variants due to LD, we found that neuronal (iN-Glut-20 or NPC-20) ASoC SNPs are highly enriched (~70-fold, p-value = 9.1×10^{-10}) for SZ risk variants, while the enrichments in OCRs or other functional annotations are much weaker (< 10-fold) (see [64] for details).

Chromosome	Position	rsID	REF count	ALT count	Binomial test p-value	ASoC FDR	SZ GWAS proxy ($R^2 \geq 0.8$)	SZ Credible SNP p-value	Genomic region	Nearest Gene	Note
chr14	99246457	rs12895055	475	164	5.42E-36	1.10E-32	rs12895055	1.08E-07	intronic	BCL11B	ASoC in NPC-20
chr1	150067621	rs2027349	122	391	7.30E-34	1.32E-30	rs2027349	1.73E-08	UTR5	VPS45(NM_007259:c.-237A>G)	ASoC in NPC-20
chr5	154258659	rs520843	16	92	3.78E-14	1.18E-11	rs520843	7.50E-08	intronic	GALNT10	ASoC in NPC-20
chr1	8408312	rs301791	69	154	1.25E-08	1.57E-06	rs301791	1.46E-08	intronic	RERE	ASoC in NPC-20
chr5	154258739	rs506674	17	66	5.60E-08	6.09E-06	rs506674	7.49E-08	intronic	GALNT10	Same region as rs520843
chr15	84570741	rs11633075	100	38	1.29E-07	1.28E-05	rs11633075	7.91E-10	ncRNA_exonic	UBE2Q2P1	
chr14	99246608	rs11624408	178	93	2.70E-07	2.49E-05	N/A	6.68E-08	intronic	BCL11B	Same region as rs12895055
chr5	140807072	rs4151680	79	149	4.14E-06	2.71E-04	rs4151680	1.67E-06	intronic	PCDHA1	
chr2	232928262	rs186132169	116	56	5.56E-06	3.50E-04	N/A	7.52E-09	intronic	NGEF	
chr1	97870773	rs9661794	161	90	8.74E-06	5.13E-04	N/A	3.40E-10	intronic	DPYD	
chr1	8409277	rs301789	56	111	2.51E-05	1.27E-03	rs301789	1.06E-08	intronic	RERE	Same region as rs301791, ASoC in NPC-20
chr14	103561933	rs7148456	101	49	2.63E-05	1.32E-03	rs7148456	2.30E-12	exonic	BAG5	
chr1	8409224	rs301790	74	132	6.45E-05	2.85E-03	rs301790	1.57E-08	intronic	RERE	Same region as rs301791, ASoC in NPC-20
chr11	130862205	rs7936858	167	245	1.42E-04	5.35E-03	N/A	2.69E-09	upstream	LOC100507431(dist=95)	
chr7	105012818	rs2192932	529	661	1.44E-04	5.40E-03	rs2192932	6.63E-09	ncRNA_exonic	KMT2E-AS1	
chr3	52685800	rs10933	157	99	3.48E-04	1.10E-02	N/A	2.85E-09	5UTR	PBRM1	
chr12	57096317	rs324015	4	21	9.11E-04	2.30E-02	rs324015	2.92E-07	UTR3	STAT6(NM_001178078:c.*255A>G)	
chr14	103855689	rs3861678	39	73	1.69E-03	3.71E-02	N/A	7.91E-12	ncRNA_intronic	LINC000637	
chr20	38812275	rs6071578	37	70	1.84E-03	3.95E-02	N/A	1.09E-08	intronic	PPP1R16B	
chr3	17803260	rs17200916	157	109	3.87E-03	6.83E-02	rs17200916	3.13E-06	intergenic	LOC105376975(dist=57598), LOC339862(dist=159312)	ASoC in NPC-20
chr3	36843901	rs9882911	40	70	5.45E-03	8.71E-02	rs9882911	6.43E-11	intronic	TRANK1	
chr15	78620601	rs7170068	29	55	6.04E-03	9.30E-02	rs7170068	N/A	intronic	CHRNA3	
chr1	8408218	rs301792	60	95	6.13E-03	9.41E-02	rs301792	1.58E-08	intronic	RERE	Same region as rs301791
chr12	57094031	rs324017	166	217	1.05E-02	1.35E-01	rs324017	2.13E-07	intronic	NAB2	
chr16	29926331	rs12716973	335	273	1.33E-02	1.55E-01	rs12716973	N/A	upstream	KCTD13	

Table 2.2: Detailed information of iN-Glut-20 heterozygous SNPs that overlap with SZ genome-wide significant SNPs and credible SNPs (chromosome positions are in hg38 coordinates).

Leveraging the functional information of ASoC SNPs and their observed relevance to SZ disease risk, we identified 20 non-MHC ASoC SNPs associated with SZ (Table 2.2) as putative SZ causal SNPs, and further assessed their regulatory potential and likely *cis*-target genes using CRISPR screening. Specifically, we used a modified CROP-seq[22] approach on NPCs (n = 3 lines) stably expressing dCas9/KRAB and repressed the regulatory activities at sequences near these ASoC SNPs using targeting gRNAs (3 gRNAs per SNP locus, Fig. 2.2). We analyzed single-cell RNA-seq of 4,099 cells, of which 2,522 were assigned to unique gRNAs (Figs. 2.5b-d). We performed differential gene expression analysis on these uniquely assigned cells, comparing cells with a certain gRNA with those in the negative control group using the edgeR-QLF test[73]. We further performed permutation tests to calibrate the inflation in test p-values that is commonly observed in DE tests on scRNA-seq data[74]. Given the relatively small number of cells per gRNA group and the limited testing power, we limited the candidate pool to *cis*-genes of these targeted sites, and used a cutoff of empirical p-value < 0.05 to identify DE genes (Fig. 2.5f). As expected from transcriptional repression by KRAB, we observed a trend towards reduced expression in identified *cis*-gene targets, with 15 out of 19 genes within 500 kb of gRNA-targeted sites showing reduced expression (p-value = 0.007, binomial test; Fig. 2.5e). In total, 10 ASoC loci were found to have *cis*-targets, 4 of which had targets shared with brain eQTL and/or brain/neural Hi-C[84]–[86], including 2 with distal targets (*APOPT1* and *LRP1*) supported by long-range Hi-C chromatin contacts[64]. Independent multiplex CRISPRi/qPCR validation in NPCs confirmed that repression at six of these ASoC loci indeed resulted in reduced expression of their corresponding *cis*-genes (Fig. 2.5g, Table 2.3).

Among these, ASoC SNP rs2027349 at the *VPS45* locus is particularly interesting. Being the most significant ASoC in iN-Glut-20 and NPC-20, it is also the only SNP with high PIP (0.45) in

its fine-mapping locus (Fig. 2.5i, see [64] for details), and targeting it resulted in a strong repressive effect on the expression of *VPS45* in CRISPRi (Fig. 2.5g). The regulatory effect of rs2027349 on VPS45 was further confirmed by CRISPR/Cas9 allelic editing of two different iPSC lines heterozygous at rs2027349 (from A/G to A/A and G/G; Fig. 2.5h). Consistent with A allele being associated with increased *VPS45* expression (Fig. 2.5j), the rs2027349-flanking OCR in isogenic CRISPR-edited A/A lines also showed higher chromatin accessibility compared with the G/G lines (Fig. 2.5h). Interestingly, rs2027349 is also within a TF-binding footprint, with its G allele predicted to disrupt the motif of EGR3/4 (Fig. 2.5h), TFs involved in neurodevelopment and SZ[87]. Taken together, our results highlight rs2027349 as a functional SZ risk variant and *VPS45*, a gene involved in vesicle-mediated protein trafficking and neurotransmitter release[88], as the likely risk gene.

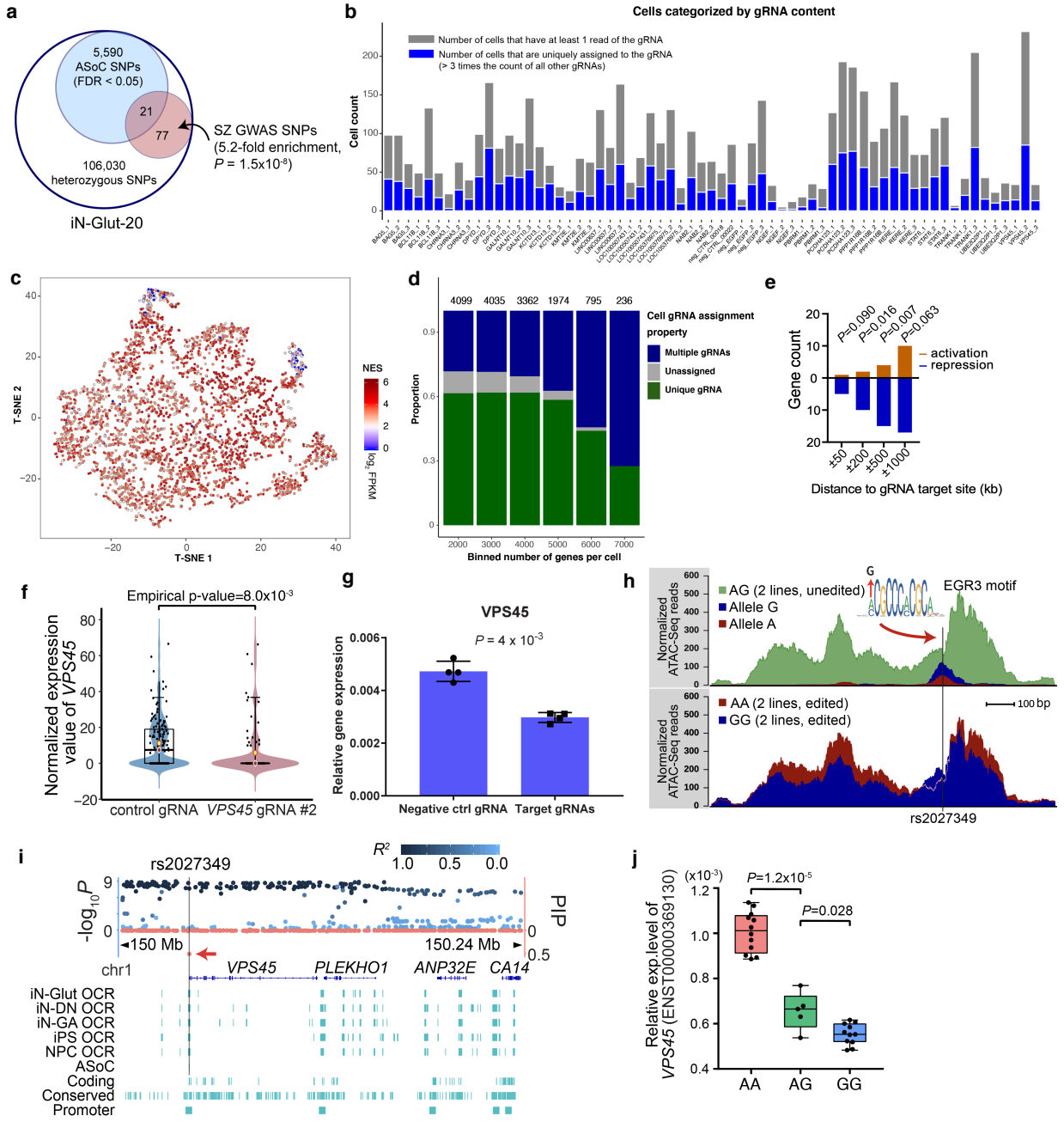

Figure 2.5: CRISPRi characterization of SZ-associated ASoC SNPs.

a) Enrichment of iN-Glut-20 ASoC SNPs for SZ GWAS SNPs. **b**) Number of cells categorized by their gRNA content, out of all cells contain a designated type of gRNAs. **c**) t-SNE plot of the 2,522 cells assigned to unique gRNAs, colored by the expression level of NES, which reflects the high purity of the NPC population used. **d**) The proportion of single cells that contained single unique gRNAs as a function of the number of expressed genes per cell. **e**) Enrichment of transcriptional repression in *cis*-genes of gRNA-targeting sites from CROP-seq. Shown are p-values of binomial enrichment test at different distance intervals. **f**) Violin/whisker plot showing reduced VPS45 expression in single NPCs with gRNA targeting rs2027349 from CROP-seq screening. **g**) Result of CRISPRi followed by qPCR validation. Relative gene expression was normalized to GAPDH in qPCR. P values were derived from Student's t-test.

(Figure 2.5 continued)
h) CRISPR allelic editing at rs2027349 alters chromatin accessibility of local OCR in isogenic NPC lines. The G allele of rs2027349 disrupts the *EGR3* motif. (The number of reads immediately around the A allele (~100 bp region) is lower presumably due to the predicted stronger TF-binding, while the broader OCR peak region in A/A line is more accessible.) **i**) Genomic location of rs2027349 (*VPS45*) locus and fine-mapping result. The left y-axis shows -$\log_{10}$ p-values from SZ GWAS (points above the x-axis), and the right y-axis shows PIPs (points below the x-axis, red). **j**) CRISPR allelic editing at rs2027349 alters *VPS45* expression in NPCs (the most abundant transcript ENST00000369130; by qPCR); expression was normalized to GAPDH.

gRNA locus (nearest gene)	***Cis*-gene**	**CROP-seq empirical p-value**	**qPCR p-value**	**Orthogonal evidence**
rs10933 (*PBRM1*)	*GNL3*	0.027	9×10^{-8}	eQTL, Hi-C
rs10933 (PBRM1)	*ALAS1*	0.024	2×10^{-3}	eQTL, Hi-C
rs2027349 (VPS45)	*ANP32E*	0.014	2×10^{-3}	eQTL, Hi-C
rs2027349 (VPS45)	*VPS45*	7.8×10^{-3}	4×10^{-3}	eQTL, Hi-C
rs2027349 (VPS45)	*SF3B4*	2.7×10^{-3}	8×10^{-3}	eQTL, Hi-C
rs2192932 (KMT2E)	*PUS7*	0.047	2×10^{-3}	Hi-C
rs7148456 (BAG5)	*APOPT1*	0.02	2×10^{-3}	eQTL, Hi-C
rs6071578 (PPP1R16B)	*FAM83D*	4.8×10^{-4}	4×10^{-3}	
rs11633075 (UBE2Q2P1)	*ZSCAN2*	0.034	0.03	eQTL

Table 2.3: SZ-associated ASoC loci with CROP-seq cis-target genes verified by qPCR assay.

2.4 Discussion

We have provided a snapshot of the ASoC landscape in an iPSC-based neurodevelopmental model and demonstrated that ASoC is a direct functional readout of noncoding risk variants for brain disorders/traits. The enrichments of neuronal ASoC SNPs for brain enhancers, TFBSs, and brain QTLs suggest mechanistic links between chromatin accessibility and gene expression. Given the strong enrichment of ASoC variants for GWAS signals of SZ and other brain disorders/traits, our study suggests that neuropsychiatric disease risk variants frequently alter chromatin accessibility and provide a useful resource for functional interpretation of noncoding disease risk variants.

Through a combination of statistical testing and experimental validation, we discovered several functional SZ risk SNPs and their *cis*-regulating genes. These discoveries warranted future functional follow-up to establish their plausible disease causality. A recent follow-up study revealed that editing at the strongest ASoC SNP (rs2027349) near *VPS45* altered the expression of three *cis*-genes, which in turn, had a compound effect on synaptic development and function in neurons, establishing a direct link between this SZ risk variant and disease-related cellular phenotypes[89].

Allelic imbalance of chromatin accessibility has proven to be a useful functional readout that is more informative than open chromatin regions. Therefore, it is crucial to maximize the mapping power of ASoC for cell types of interest. After evaluating how the sensitivity of detecting ASoC varies with SNP effect sizes and read depth, we concluded that our study has excellent power to detect ASoC with moderate to strong OCR signal (reads > 100; our median read-depth = 59). And by including those OCRs and SNPs with read depth of 20-100, we expected to capture ASoC SNPs with relatively weak OCRs but strong SNP effect sizes. However, due to the relatively small sample size, our power of ASoC mapping is limited for SNPs with low coverage and effect sizes, especially for cell types with only 8 sequenced samples. This limited the types of comparisons that can be made between cell types, which could potentially shed light on the key regulatory mechanisms differentiating various neuronal subtypes. Deep sequencing of samples rich in heterozygous SNPs (that target genome regions near important genes) may improve the power of ASoC mapping.

Prioritizing functional disease risk variants and pinpointing causal variants has always been a challenge because a GWAS locus often contains numerous common SNPs associated with the trait due to LD. In our study, we prioritized non-coding disease variants by intersecting our neuronal

ASoCs with GWAS SZ index SNPs. Although statistical fine-mapping that leverages ASoC information was later used to interpret our findings from CRISPRi screening, it could be applied at an earlier stage to further narrow down the putative GWAS causal SNPs. This approach can be utilized to relieve the burden for experimental screening, especially when there remain many candidates.

We assessed the regulatory functions of tagged sequences of putative SZ causal SNPs using multiplexed CRISPRi screen followed by single-cell RNA-sequencing (CROP-seq), which has the benefit of higher throughput and resolution compared with traditional CRISPR screen assays. To identify affected genes under each perturbation based on the transcriptomic readout of single cells, we experimented with several routine differential expression methods in addition to the adopted edgeR-QLF test, including the Wilcoxon rank sum test, edgeR-LRT test[73], DESeq2[90], and MAST[91]. However, we found that routine DE approaches tend to be under-powered when applied to single-cell screening data, which limited us to only consider *cis*-genes of targeted regions and use a relatively lenient cutoff on test statistics without multiple correction. We believe this lack of detection power, especially for *trans*-effects, is due to several reasons. First of all, as CRISPRi interference merely repressed the activity of putative regulatory regions instead of introducing gene-level knockouts, the effect sizes imparted are relatively small. Likewise, *trans*-regulatory effects of genetic variants, being mediated through genes, are generally small in effect size as well, as reflected by the difficulties in *trans*-eQTL mapping[92]. Secondly, because of low gRNA efficiency, relatively small numbers of cells (in our case < 100) are assigned to a gRNA. Therefore, one needs to increase the number of single cells in the experiment or use a high MOI design[23] to increase the sample size for each gRNA group. Thirdly, compared with bulk RNA-seq, scRNA-seq data are sparse and noisy in nature, and require careful data normalization and

statistical modeling with proper assumptions that can better uncover the true signals without introducing artifacts to the data[29], [93], [94]. Given these challenges and the increasing interest of the research community in single-cell CRISPR screen, novel statistical analysis appropriate for this type of data will be extremely useful. This prompted the third chapter of my PhD research.

CHAPTER 3: A NOVEL BAYESIAN FACTOR ANALYSIS METHOD FOR SINGLE-CELL CRISPR SCREENING DATA[1]

3.1 Introduction

In this chapter, we continue the studying of genetic variations, focusing on understanding their effects on the transcriptome through single-cell RNA-sequencing (scRNA-seq) combined with CRISPR screening. As introduced in Chapter 2, the CRISPR-Cas9 genome engineering system is a powerful tool that can manually introduce genetic perturbations and screen for genes of important biological functions. Technologies such as CROP-seq[22] and Perturb-seq[21] have further improved the power of CRISPR by combining multiplexed CRISPR screening with scRNA-seq. By linking guide RNAs (gRNAs) to single cell transcriptomes, these technologies provide high-content molecular readouts of the target perturbations within single cells. Single-cell CRISPR screening technologies have found many applications, for example, in studies of developmental regulators, genes involved in immune responses, and regulatory elements involved in human diseases[23]–[25], [95].

Nevertheless, the analysis of single-cell CRISPR screening data remains challenging. A routine strategy for analyzing gene expression data is differential expression analysis, whereby the effects of genetic perturbation on the transcriptome are assessed one gene at a time[73], [90]. However, as we have observed in Chapter 2, when applied to single cell screening data, this type of analysis can be under-powered due to the sparsity and noise inherent to scRNA-seq data, as well as the relatively small numbers of cells (often hundreds) per perturbation in typical experiments.

[1] Much of this chapter contains material from the preprint: Zhou *et al.*, "A novel Bayesian factor analysis method improves detection of genes and biological processes affected by perturbations in single-cell CRISPR screening," *bioRxiv*, doi: https://doi.org/10.1101/2022.02.13.480282.

Another commonly used analysis is to cluster cells based on their transcriptome similarity, and then assess whether cells with a specific perturbation are enriched or depleted in any cluster[95], [96]. However, this clustering-based approach has several limitations. CROP-seq studies often use relatively homogenous populations of cells to minimize variation across cells and increase the power of discovering transcriptional effects; thus, it can be challenging to partition cells into distinct clusters. Furthermore, the clustering approach does not explicitly link the perturbations with the affected genes, limiting our understanding of the downstream effects of the perturbations. Given the limitations of standard differential expression and clustering-based analyses, rigorous statistical methods that accommodate the unique features and complexities of single-cell CRISPR screening data are greatly needed.

We propose to infer gene modules from scRNA-seq data, and borrow information across genes to improve the power of detecting differentially expressed genes. This proposed approach is motivated by the observation that genetic perturbations typically affect expression, not one gene at a time, but many related genes simultaneously. Indeed, transcriptomes often vary across cell types, conditions, and individuals in a modular fashion, reflecting the underlying coordinated regulation of genes in the same or related pathways. These gene modules can be inferred by matrix factorization and related techniques (introduced in Section 3.2). Existing factor analysis methods, however, are not readily applied to single-cell CRISPR screening data, as the factors are not directly linked with genetic perturbation, and the effects of perturbation on the expression of individual genes are not assessed (more discussion in Section 3.5).

In Section 3.3, we present Guided Sparse Factor Analysis (GSFA), a statistical framework for analyzing single-cell CRISPR screening data that bridges factor analysis and differential expression analysis. GSFA assumes the effects of genetic perturbations are mediated through a set

of gene modules representing biological pathways or functional units, and captures them mathematically as latent factors. GSFA evaluates associations of the genetic perturbations with these latent factors, providing information on the module-level effects of the perturbations. Compared with single-gene level differential expression analysis, this factor association analysis may be more sensitive. Indeed, expression of a single gene is noisy, and is influenced by potentially many sources. In contrast, latent factors represent the main dimensions of variation of many genes, and can be thought of as "denoised" versions of gene expression, possibly improving the detection of the effects of perturbations. While our approach is formulated in terms of latent factors, we can still summarize the effects of a perturbation on individual genes as the sum of effects mediated by all the factors. We benchmarked our method through extensive simulation studies and real data applications in Section 3.4. Overall, GSFA identifies biologically relevant modules, and has better power to detect differentially expressed genes (DEGs) than alternative methods, providing novel insights into the molecular mechanisms of important biological processes such as regulation of T cell activation and neuronal differentiation.

3.2 Matrix Factorization Methods for High-Dimensional Genomics Data

High-throughput omics technologies have generated an unprecedented amount of high-dimensional data that record measurements of features such as gene expression, methylation, and protein concentration levels in individual samples. While these high-dimensional datasets provide researchers with comprehensive measures of cellular response, they also demand statistical methods that can effectively analyze and interpret the data. Oftentimes, samples in these datasets are not independent, but instead share similar phenotypes based on their tissue types, cell types, experimental replicates, etc. Patterns of relatedness are also prevalent among biological features, for example, genes in the same biological pathway tend to have co-regulated expression levels.

These modular structures make it possible to represent high-dimensional omics data in a lower dimensional subspace via statistical methods. One class of unsupervised methods, matrix factorization, can perform such mappings linearly and learn the lower dimension structures from the original data[97]. For a typical omics data matrix **Y** that holds the measurements across N samples and P features, matrix factorization methods learn an $N \times K$ matrix **Z** that describes the relationship between samples, and a $P \times K$ matrix **W** that describes the relationship between features, with K here being the number of factors, or the rank or dimensionality of the matrix factorization problem. **E** is an error term matrix that holds the residual errors that cannot be explained by factors:

$$\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{W}^{\boldsymbol{T}} + \boldsymbol{E}. \quad (1)$$

In the context of transcriptomics data where Y_{ij} represents the expression level of gene j in sample i, each factor usually represents a set of genes with coordinated expression levels, *i.e.*, a gene module. We refer to **W** as the gene loading matrix. The values in each column of **W** are gene weights that reflect the contribution of genes in each factor, and can be used to associate the factor with biological processes or functional pathways under certain model specifications. **Z** is a lower dimensional representation of the samples, with each row describing how active the factors are in each sample, therefore, we call **Z** the factor loading matrix. Through methods such as clustering analysis on **Z**, the relationships among samples can be learned.

Numerous matrix factorization methods have been applied to omics data, assuming different properties on the factor and loading matrices to facilitate the downstream interpretation of different biological problems. Here, we attempt to provide an overview of some major classes of matrix factorization methods for genomics data.

3.2.1 Overview of matrix factorization methods

Principal components analysis (PCA) is a popular matrix factorization method that finds a set of orthogonal principal components (PCs) as the projection of the original data, with each PC capturing the maximal variance possible[98]. By only keeping the first few principal components, it can reduce dimensionality of the original data while preserving as much variation as possible. The top PCs can be computed via truncated singular value decomposition (SVD) of the observed data matrix. Assuming that **Y** is column-centered, a rank-*K* truncated SVD of **Y** is

$$\boldsymbol{Y} = \boldsymbol{UDV}^T. \quad (2)$$

Then the columns of **UD** are equivalent to the top *K* PCs, while the columns of **V** are the *K* eigenvectors of the covariance matrix of **Y** that correspond to its *K* largest eigenvalues, and can be interpreted as the feature loadings to PCs. Relating this to matrix factorization, PCA can be interpreted as the least square solution for Equation (1) under the constraints that columns of **Z** are orthogonal and columns of **W** are orthonormal.

PCA has been widely used for processing and interpreting high-dimensional genomics data, such as exploratory analysis for latent structures (*e.g.* clusters) among samples, and identification of genes that distinguish between samples. By associating the inferred PCs with observed covariates of the samples, one can identify key sources that influence the variation in data. However, PCA is conceptually different from factor analysis. By construction, the PCs are orthogonal components that maximize variability, and therefore, are determined by the observed variables, while factor analysis finds latent factors that cause the responses in observed variables. The results of PCA also lack in interpretability as the loadings produced are usually dense, which means the PCs do not correspond to individual biological processes, but instead, are mixtures of multiple processes[97].

Non-negative matrix factorization (NMF) is another group of methods that has found wide application in genomics data, for example, in sample clustering and detection of genes that differentiate sample groups. It is particularly suited for interpreting transcriptome and methylome data[99], [100], as it requires the input data to be non-negative, and finds a non-negative matrix factorization of rank *K* that best approximates the input matrix. Typically, a loss function of the Frobenius norm (Equation 3) is used, and NMF is equivalent to fitting (1) with normal errors under the constraints that both the factor and feature loading matrices are non-negative, although alternative formulations of NMF exist[101].

$$\min_{\boldsymbol{Z},\boldsymbol{W}} \|\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{W}^T\|_F^2 \;\; s.t. \boldsymbol{Z} \geq \boldsymbol{0}, \boldsymbol{W} \geq \boldsymbol{0}. \tag{3}$$

Since each observed sample (a row of **Y**) is now approximated by a linear combination of components with non-negative weights given by each row of **Z**, this is compatible with the intuitive notion that an addition of parts form an entirety[102]. Therefore, in contrast to PCA, where the components are ranked solely by the variance they explained, NMF tends to learn equally important biological processes that make up the observed data. Despite being more intuitive, however, the NMF problem is computationally hard as the factorization is not unique. The non-negative constraint on input also makes NMF less flexible than other matrix factorization models, and the construction of non-negative gene weights limits the identification of genes to only those that are overexpressed in components, but not the ones that are repressed[97].

Factor analysis (FA) is a class of statistical methods that infers low-rank components from high-dimensional data using a generative model with specific assumptions on the probabilistic distributions of its components. Extending on Equation (1), classic FA assumes a generative model where the rows of **Y** and **Z** are independently and identically distributed (Equation 4). The marginal distributions of elements in **Z** are assumed to be standard normal, and the conditional

distribution of each row of **Y** is a multivariate Gaussian with diagonal residual covariance that allows for column-wise variances[103]. Therefore, FA learns the latent factors that drive the observe variable response, and given these factors, the observed variables are independent.

$$y_i \sim N(\boldsymbol{W}z_i, \boldsymbol{\Sigma}), \boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \dots, \sigma_P^2), z_i \sim N(0, \mathbf{I}_K). \quad (4)$$

As with NMF, classic FA faces the challenge of unidentifiability, in that factor and gene loading matrices are unique only up to an orthogonal transformation. And similar to PCA, the loadings inferred in FA are mostly nonzero, making it difficult to gain biological interpretation. Just like how variable selection can help obtain identifiable and parsimonious models for high-dimensional linear regression problems, these issues in matrix factorization can be mitigated in the same way by introducing sparsity on the factor or gene loading matrices, such that only the most representative genes have nonzero loadings to factors, and each sample consists of only a few factors. In practice, sparsity can be induced by adding regularizations or priors to **Z** and **W**, giving rise to frequentist or Bayesian variations of sparse factorization methods based on PCA, NMF or FA. The collective of these methods have not only extended the knowledge on factor model solutions, but also contributed substantially to the interpretation of complex genomics data.

3.2.2 Sparse matrix factorization with regularization

Regularization using penalties such as the l_1 norm (lasso) is a frequentist variable selection approach to achieve model parsimony. Similar to penalized regression, sparse PCA[104] can obtain PCs with sparse loadings by writing PCA in a regression-type optimization problem, and subjecting the feature loadings (**B**) to elastic net penalty terms that allow for different tuning parameters per component (Equation 5).

$$\left(\widehat{\boldsymbol{A}}_{P\times K}, \widehat{\boldsymbol{B}}_{P\times K}\right) = \underset{A,B}{\text{argmin}} \sum_{i=1}^{N} \|y_i - \boldsymbol{A}\boldsymbol{B}^T y_i\|^2 + \lambda \sum_{k=1}^{K} \|b_k\|^2 + \sum_{k=1}^{K} \lambda_{1,k} \|b_k\|_1$$
$$\text{subject to } \boldsymbol{A}^T\boldsymbol{A} = \mathbf{I}_K. \quad (5)$$

Sparse SVD can be achieved by adding adaptive lasso penalty terms that contain data-driven weights on elements in both the left and right singular vectors, which results in the biclustering of both samples and features[105]. The regularized NMF framework suggested in[106] accommodates for a variety of penalty terms on one of the lower-rank matrices while fixing the scale of the other unregularized matrix. In addition to detecting the underlying modular structures, applications of these regularized methods on gene expression data better capture lower-rank components that can be biologically annotated using methods such as gene set enrichment analysis[99], [106]. PLIER (Pathway-level information extractor)[107], another method built upon PCA, further improved the interpretability of PCs by directly incorporating existing biological knowledge into the constraints on gene loadings such that they align as much as possible with, for example, a series of gene sets that represent known biological pathways.

However, the increase in interpretability comes with a cost. The addition of penalty terms constrains the problem to an optimization one, where iterative algorithms such as coordinate descent are typically needed to find the solutions. This optimization problem becomes even more challenging when non-convex penalties (*e.g.* l_q pseudo-norm, $0 < q < 1$) are involved. Moreover, since there lack analytical solutions for the regularization tuning parameters, their values have to be determined empirically through model evaluation methods such as cross validation[108]. Hence, figuring out the best combination of tuning parameter settings can be computationally intensive, especially when multiple parameters are involved.

3.2.3 Bayesian sparse factor analysis

Model sparsity can also be achieved in the Bayesian framework using shrinkage priors, many of which have been proven to have performance comparable to or better than their frequentist counterparts in linear regression[109], [110]. By specifying a sparse prior distribution on the model

parameters, the Bayesian approach fits naturally in with generative models such as factor analysis, shrinking the elements in loading matrices toward zero. A simple form of shrinkage is assuming a normal prior on the parameters of interest, the posterior mean of which is effectively the estimate for ridge regression[111]. When applied to the factor loadings in factor analysis (Equation 6), this leads to strong shrinkage of elements in $\mathbf{Z}$ toward 0 wherever it is consistent with the data, producing more interpretable results than PCA[112].

$$Y_{ij} \sim N((\boldsymbol{ZW}^T)_{ij}, \psi_i^{-1}),\ Z_{ik} \sim N(0, \sigma_{ik}^2) \quad (6)$$

Another category of commonly used shrinkage priors are normal-mixture related priors. An initial form proposed in [113] is in fact a "spike-and-slab" prior, which models a variable as coming from a discrete mixture of a point mass at zero (the spike) and a normal distribution that is also centered at 0 (the slab) (Equation 7). The mixture proportion for the normal distribution, π, can be interpreted as the prior probability that the variable is non-zero, or that the corresponding predictor should be included in the mode. Therefore, the estimate of 1-π reflects the sparsity level of the final model.

$$g(\cdot) = \pi N(\cdot; 0, \sigma_1^2) + (1 - \pi)\delta_0(\cdot) \quad (7)$$

When imposed on the gene loading matrix in gene expression data, for example, the spike-and-slab prior shrinks the loadings of genes with small effects to exact zeros, resulting in more interpretable factors that are only linked to a number of genes[114]. Placing spike-and-slab priors on elements in both the factor and feature loading matrices allows for biclustering of both samples and features, which has proved useful in data-driven detection of linear structure present in gene expression and other sources of data[115]. The "normal-mixture" prior[116] is a continuous version close to the spike-and-slab prior. It consists of two normal densities centered at 0 but with different variances (Equation 8), where one of the variances is a smaller number that serves similar

effects as the spike in inducing sparsity in data, pulling small effects close to but not exactly 0, while preserving larger effects, offering more flexibility than the spike-and-slab prior:

$$g(\cdot) = \pi N(\cdot; 0, \sigma_1^2) + (1 - \pi) N(\cdot; 0, \sigma_2^2). \quad (8)$$

Shrinkage priors that are even more flexible than the two-component normal-mixture prior also exists. For example, the adaptive shrinkage ("ash") prior, which is a unimodal prior that assumes a mixture of point mass at zero and normal distributions with zero means (Equation 9), can not only adapt for the sparsity level, but also capture the tail behavior of non-zero effects in the data through the grid of variances[117]. Application of the ash prior in factor analysis allows the model to adapt to any combinations of sparsity in the factor and feature loadings[118], outperforming other frequentist approaches in structure detection and factor interpretability.

$$g(\cdot) = \pi_0 \delta_0(\cdot) + \textstyle\sum_k \pi_k N(\cdot; 0, \sigma_k^2). \quad (9)$$

Under the Bayesian framework, parameters in these sparse priors can be estimated together with other model parameters, either using "full Bayes" solutions[59], [114] or via empirical Bayes methods[118], [119]. In other words, parameters that control shrinkage in sparse factor models are learnt from data and estimated by sampling from the posterior distributions, while frequentist regularization approaches rely on cross validation to determine the penalty parameters. For this reason, it is also easy to obtain the uncertainty of these parameters from the posterior samples. In addition, the posterior estimates often have intuitive interpretations, for example, the posterior inclusion probability (PIP) directly reflects how likely a variable is included in the true model given the observations. Finally, in contrast to the optimization problems in the frequentist framework, the inference of Bayesian model parameters usually relies on Markov chain Monte Carlo (MCMC)[120], the implementation of which is relatively easy regardless of whether the

shrinkage priors correspond to convex or non-convex penalties, providing more flexibility in the choice of priors[121].

Choosing the optimal latent dimensionality (K) has always been a challenge in matrix factorization problems[97]. Although model selection can be used to compare models of different dimensionalities, it still requires a specific range of K's to search over manually. Under the Bayesian framework, this matter can be addressed by nonparametric modeling of the latent structure or the introduction of infinite number of factors followed by increased shrinkage on their loadings[114], [122], [123].

Overall, Bayesian sparse factor analysis with its model flexibility and straightforward implementation, has found wide application in the interpretation of high-dimensional genomics data. Various forms of sparse priors can be applied to accommodate different latent structures underlying the data. One needs to carefully consider the dataset in hand and the questions in mind when choosing the appropriate factorization method, as they each have strengths and weaknesses under distinct biological contexts.

3.3 Guided Sparse Factor Analysis (GSFA)

Here, we describe GSFA, a Bayesian sparse factor analysis model tailored for analyzing single-cell CRISPR screening data that unifies factor analysis and estimation of the effects of target perturbations. GSFA assumes that the perturbation of a target gene affects certain latent factors, which in turn changes the expression of individual genes. In addition to the sparse priors imposed on the gene loadings, GSFA adds a second layer to the factor analysis model that links the factors to the perturbations via a multivariate linear regression. In this way, GSFA can detect interpretable gene modules and DEGs affected by genetic perturbations for a large number of cells and multiple perturbations simultaneously. GSFA can also accommodate different cell types or multiple

experimental conditions, where it estimates the effects of perturbation separately and produce different DEGs for each cell type/condition. GSFA is implemented in R with Rcpp acceleration, and distributed as a freely available R package at Github: https://github.com/gradonion/GSFA.

3.3.1 The GSFA model

The input of GSFA consists of two matrices: a normalized gene expression matrix $\mathbf{Y}_{N\times P}$ with N cells and P genes, and a "perturbation matrix" $\mathbf{G}_{N\times M}$ that records M types of genetic perturbations in those N cells. In the simplest case, the perturbation matrix is binary, *i.e.*, $\mathbf{G}_{ij}$ is 1 if cell i has the m-th type of perturbation and 0 otherwise, but this is not strictly required by the model, *e.g.*, $\mathbf{G}$ might represent the dosage of genetic perturbations. Under the assumptions mentioned above, the GSFA model has two main parts: (1) a sparse factor analysis model that decomposes the expression matrix $\mathbf{Y}$ into a factor matrix $\mathbf{Z}_{N\times K}$, where K is the number of factors, and a sparse gene loading matrix $\mathbf{W}_{P\times K}$, and (2) a multivariate linear model that captures the dependency of factors ($\mathbf{Z}$) with the perturbation matrix $\mathbf{G}$, which makes our model is a form of "guided" factor analysis (Fig. 3.1):

$$\mathbf{Y} = \mathbf{Z}\mathbf{W}^T + \mathbf{E}, E_{ij} \sim N\left(0, \psi_j\right), \quad (10)$$

$$\mathbf{Z} = \mathbf{G}\boldsymbol{\beta} + \boldsymbol{\Phi}, \phi_{ik} \overset{i.i.d}{\sim} N(0,1) \quad (11)$$

Here $\mathbf{E}$ is an $N\times P$ residual matrix with gene-specific variances stored in a P-vector $\boldsymbol{\psi}$, $\boldsymbol{\beta}$ is an $M\times K$ matrix of perturbation effects on factors, and $\boldsymbol{\Phi}$ is an $N\times M$ residual matrix with variance 1.

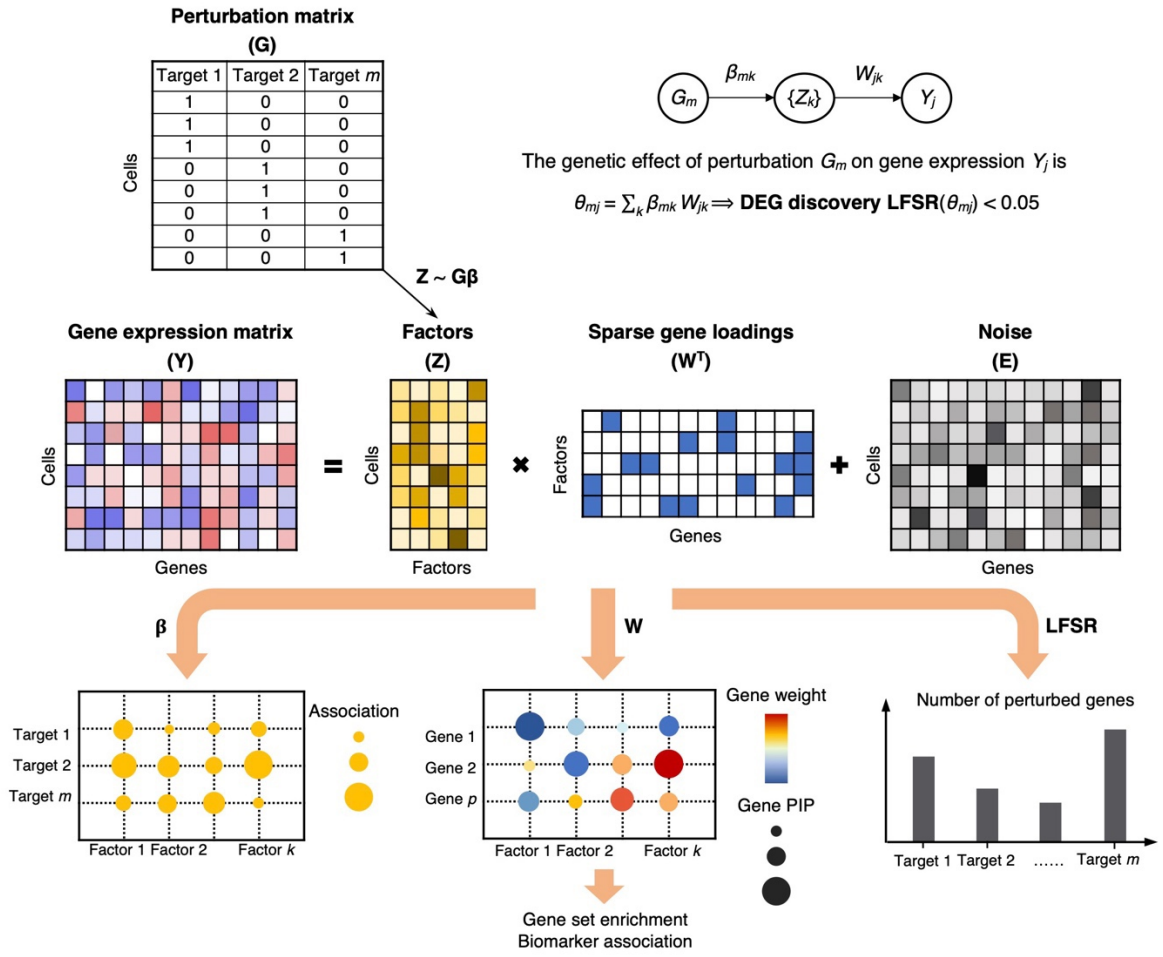

Figure 3.1: Schematic of the GSFA model and its application on real data.
Top: the input of GSFA includes the perturbation matrix and the gene expression matrix; bottom: the output of GSFA includes the effects of perturbations on factors (**β**), the gene loading matrix (**W**), and the list of genes affected by each perturbation after LFSR thresholding.

We assume that each genetic perturbation likely affects only a small number of factors, and impose a spike-and-slab prior on the effect of perturbation m ($1 \leqslant m \leqslant M$) on factor k ($1 \leqslant k \leqslant K$):

$$\beta_{mk} \sim p_m N(0, d_m^2) + (1 - p_m)\delta_0, \qquad (12)$$

where δ_0 is delta-function, p_m denotes the proportion of factors affected by perturbation m, and d_m the prior variance of the effect sizes of perturbation m.

To limit the number of genes contributing to a factor and facilitate the biological interpretation of factors, we also impose a sparse prior on the gene loading matrix **W**. We evaluated two choices,

the standard spike-and-slab prior, and a normal-mixture prior[116], where the effect is sampled from a mixture of two normal distributions, one "foreground" component capturing true effects, and the other a "background" component absorbing small effects that is not necessarily a point mass at 0. This comparison is motivated by a well-known problem in count-based RNA-seq data analysis: because the total read count in a sample is fixed, activation of some genes indirectly reduces the read counts in all other genes, resulting in weakly correlated expression across many genes. Thus, even when a factor affects only a small set of genes, it may appear to be correlated with many other genes, making it hard to infer sparse factors. A normal-mixture prior may help address this problem, because of its "background" component with small effects. Indeed, it shows better results in our simulations compared with the spike-and-slab prior (see Section 3.4.1), so is used as our default prior. Specifically, the prior weight of gene *j* in the factor *k* follows:

$$W_{jk} \sim \pi_k N(0, \sigma_k^2) + (1 - \pi_k) N(0, \sigma_k^2 c_k^2), 0 < c_k < 1 \quad (13)$$

where π_k represents the proportion of genes affected by the factor *k* (the "foreground" part), and σ_k^2 the prior effect size variance of factor *k*, and c_k a scale parameter controlling the relative size of the foreground and background effects.

We also specified conjugate prior distributions for other parameters in the model as follows:

$$\psi_j^{-1} \sim \text{Gamma}(g_0, h_0),$$

$$\pi_k \sim \text{Beta}(s_w r_w, s_w(1 - r_w)),$$

$$\sigma_k^{-2} \sim \text{Gamma}(g_w, h_w),$$

$$c_k^{-2} \sim \text{Gamma}(g_c, h_c),$$

$$p_m \sim \text{Beta}(s_b r_b, s_b(1 - r_b)),$$

$$d_m^{-2} \sim \text{Gamma}(g_b, h_b).$$

In practice, we set the hyperparameters of priors for the variance parameters to $g_0 = 1$, $h_0 = 1$, $g_w = 1$, $h_w = 1$, $g_c = 3$, $h_c = 0.5$, $g_b = 1$, $h_b = 1$, such that their prior means are all 1 other than $\bar{c}_k^2 = \frac{1}{6}$. For parameters that reflect the sparsity levels, π_k and p_m, tuning the hyperparameters in their prior distributions may influence the estimated sparsity outcomes of the model to various extents, and should be set accordingly based on the dimensionality of data and the amount of shrinkage one expects. We set $s_w = 50$ and $r_w = 0.2$ when the number of genes is 6000, so that the prior mean for π_k is 0.2. r_b is set to 0.2 so that the prior mean for p_m is 0.2; in simulations where we have 10 factors, $s_b = 5$, and in real data applications where 20 factors are specified, $s_b = 20$.

3.3.2 Full Bayesian inference using Gibbs sampling

We infer the parameters in GSFA using Gibbs sampling, an MCMC algorithm that obtains a sequence of approximate samples from their posterior distribution given the observed data. Gibbs sampling is an attractive choice here because the conditional distributions of the main parameters ($\boldsymbol{\beta}$, $\mathbf{W}$, and $\mathbf{Z}$) all have analytic forms, so do the hyperparameters with conjugate priors.

To see this, we first consider the conditional distribution of $\mathbf{W}$ given data and all other parameters/variables. For simplicity, we drop the hyperparameters and parameters related to the error terms. It's easy to see that given $\mathbf{Z}$, the conditional distribution of $\mathbf{W}$ does not depend on $\mathbf{G}$ and $\boldsymbol{\beta}$: $P(\mathbf{W}|\mathbf{Y}, \mathbf{G}, \mathbf{Z}, \boldsymbol{\beta}) = P(\mathbf{W}|\mathbf{Y}, \mathbf{Z})$. The problem now becomes one of multivariate linear regression, $\mathbf{Y}=\mathbf{Z}\mathbf{W}^{\mathrm{T}}+\mathbf{E}$, where $\mathbf{W}$ follows a spike-and-slab prior. This is a well-studied problem in the statistics literature[124], [125]. Similarly, we can see that the conditional distribution of $\boldsymbol{\beta}$ is given by: $P(\boldsymbol{\beta}|\mathbf{Y}, \mathbf{G}, \mathbf{Z}, \mathbf{W}) = P(\boldsymbol{\beta}|\mathbf{G}, \mathbf{Z})$. This reduces to a regression problem $\mathbf{Z} = \mathbf{G}\boldsymbol{\beta} + \boldsymbol{\Phi}$, where $\boldsymbol{\beta}$ follows a normal-mixture prior. Finally, the conditional distribution of $\mathbf{Z}$ is given by:

$$P(\mathbf{Z}|\mathbf{Y}, \mathbf{G}, \mathbf{W}, \boldsymbol{\beta}) \propto P(\mathbf{Z}|\mathbf{G}, \boldsymbol{\beta}) \cdot P(\mathbf{Y}|\mathbf{Z}, \mathbf{W}).$$

This is another regression problem $\mathbf{Y} = \mathbf{Z}\mathbf{W}^{\mathrm{T}} + \mathbf{E}$, where each row of $\mathbf{Z}$ represents the unknown linear coefficients with a normal prior, $Z_i \sim \mathrm{N}(G_i\boldsymbol{\beta}, \mathbf{I}_K)$, for sample i ($1 \leqslant i \leqslant N$).

To facilitate computation, we also introduced two latent binary matrices, $\mathbf{F}_{P\times K}$ and $\boldsymbol{\gamma}_{M\times K}$, to indicate which distributions the corresponding parameters in $\mathbf{W}$ and $\boldsymbol{\beta}$ come from. The joint prior distribution of $\mathbf{W}$ and $\mathbf{F}$ follows:

$$P\left(W_{jk}|F_{jk}=1\right)P\left(F_{jk}=1\right) = N\left(W_{jk};0,\sigma_k^2\right)\cdot\pi_k,$$

$$P\left(W_{jk}|F_{jk}=0\right)P\left(F_{jk}=0\right) = N\left(W_{jk};0,\sigma_k^2c_k^2\right)\cdot(1-\pi_k).$$

The joint prior distribution of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ can be written as:

$$P(\beta_{mk}|\gamma_{mk}=1)P(\gamma_{mk}=1) = N(\beta_{mk};0,d_m^2)\cdot p_m,$$

$$P(\beta_{mk}|\gamma_{mk}=0)P(\gamma_{mk}=0) = \delta_0(\beta_{mk})\cdot(1-p_m).$$

Now we describe the detailed Gibbs sampling update rules in GSFA. To obtain posterior samples for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, we update β_{mk} and γ_{mk} ($1 \leqslant m \leqslant M$, $1 \leqslant k \leqslant K$) in pairs. We first sample γ_{mk} from a Bernoulli distribution where the success rate is determined by the product of two ratios: the ratio of two marginal likelihoods, and the prior ratio:

$$\frac{P(\gamma_{mk}=1|\cdot)}{P(\gamma_{mk}=0|\cdot)} = \frac{P(\mathbf{Z}|\gamma_{mk}=1,\mathbf{G},\boldsymbol{\beta}_{-mk},\boldsymbol{\gamma}_{-mk},\boldsymbol{d}^2)}{P(\mathbf{Z}|\gamma_{mk}=0,\mathbf{G},\boldsymbol{\beta}_{-mk},\boldsymbol{\gamma}_{-mk},\boldsymbol{d}^2)}\cdot\frac{P(\gamma_{mk}=1|p_m)}{P(\gamma_{mk}=0|p_m)}$$

$$= \sqrt{\frac{L_{mk}}{d_m^2}}\exp\left(\frac{\mu_{mk}^2}{2L_{mk}}\right)\cdot\frac{p_m}{1-p_m}.$$

Here $\mu_{mk} = L_{mk}\sum_{i=1}^{N} G_{im}(Z_{ik} - \sum_{l:l\neq m} G_{il}\beta_{lk})$ and $L_{mk} = (\sum_{i=1}^{N} G_{im}^2 + d_m^{-2})^{-1}$.

After γ_{mk} is updated, we then sample β_{mk} from

$$\beta_{mk}|\gamma_{mk}=1 \sim N(\mu_{mk},L_{mk}),$$

$$\beta_{mk}|\gamma_{mk}=0 \sim \delta_0.$$

For the remaining parameters, we draw their posterior samples from the following posterior distributions:

$W_{j\cdot}|\cdot \sim N(\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}+\mathbf{D}_j\right)^{-1}\mathbf{Z}^T Y_{\cdot j},\ \psi_j\left(\mathbf{Z}^{\mathrm{T}}\mathbf{Z}+\mathbf{D}_j\right)^{-1})$,

where $\mathbf{D}_j = \mathrm{diag}(\frac{\psi_j}{\sigma_1^2[F_{j1}+(1-F_{j1})c_1^2]}, \dots, \frac{\psi_j}{\sigma_K^2[F_{jK}+(1-F_{jK})c_K^2]})$,

$F_{jk}|\cdot \sim \mathrm{Bernoulli}(\frac{r_{jk}}{r_{jk}+1})$, where $r_{jk} = \frac{\pi_k}{1-\pi_k}c_k \exp\left(\frac{W_{jk}^2}{2\sigma_k^2}(\frac{1}{c_k^2}-1)\right)$,

$Z_{i\cdot}|\cdot \sim N(\mu_i, \Sigma)$,

where $\mu_i = \Sigma\cdot(\mathbf{W}^{\mathrm{T}}\mathbf{\Psi}^{-1}Y_{i\cdot}+\boldsymbol{\beta}G_i)$, $\Sigma = (\mathbf{W}^{\mathrm{T}}\mathbf{\Psi}^{-1}\mathbf{W}+\mathbf{I}_K)^{-1}$, and $\mathbf{\Psi} = \mathrm{diag}(\psi_1, \dots, \psi_P)$.

$\psi_j|\cdot \sim \mathrm{InverseGamma}(g_0+\frac{N}{2}, h_0+\frac{1}{2}\sum_{i=1}^N\left(Y_{ij}-\sum_{k=1}^K Z_{ik}W_{jk}\right)^2)$,

$\pi_k|\cdot \sim \mathrm{Beta}(s_w r_w+\sum_{j=1}^P F_{jk}, s_w(1-r_w)+P-\sum_{j=1}^P F_{jk})$,

$\sigma_k^2|\cdot \sim \mathrm{InverseGamma}(g_w+\frac{P}{2}, h_w+\frac{1}{2}\sum_{j=1}^P \frac{W_{jk}^2}{F_{jk}+(1-F_{jk})c_k^2})$,

$c_k^2|\cdot \sim \mathrm{InverseGamma}(g_c+\frac{1}{2}\sum_{j=1}^P(1-F_{jk}), h_c+\frac{1}{2}\sum_{j:F_{jk}=0}\frac{W_{jk}^2}{\sigma_k^2})$,

$p_m|\cdot \sim \mathrm{Beta}(s_b r_b+\sum_{k=1}^K \gamma_{mk}, s_b(1-r_b)+K-\sum_{k=1}^K\gamma_{mk})$,

$d_m^2|\cdot \sim \mathrm{InverseGamma}(g_b+\frac{1}{2}\sum_{k=1}^K\gamma_{mk}, h_b+\frac{1}{2}\sum_{k=1}^K\beta_{mk}^2)$.

The computational complexity of the Gibbs inference is $O(N(P+M)K)$ per iteration, with N being the number of cells, P being the number of genes, M being the number of perturbations, and K being the number of factors. The average run time on a simulated dataset of 4000 cells, 6000 genes, and 10 factors is 1.32 seconds per iteration on a modern Linux workstation with Intel Xeon E5-2680 v4 (2.40 GHz) processors.

3.3.3 Posterior inclusion probabilities and summary of perturbation effects on individual genes

For any parameter with sparse prior, the probability that its value is non-zero or comes from the foreground component, is denoted as the posterior inclusion probability (PIP). PIPs can be easily obtained from the posterior distribution. The PIP of β_{mk} quantifies whether a perturbation affects a certain factor, and can be computed from the posterior mean of γ_{mk}:

$$\mathrm{PIP}(\beta_{mk}) := \Pr(\beta_{mk} \neq 0 | \mathrm{Data}) = \Pr(\gamma_{mk} = 1 | \mathrm{Data}).$$

The PIP of W_{jk} quantifies whether a gene has loading on a factor, which, in our case, also means the probability of W_{jk} coming from the "foreground" normal distribution given data, and can be computed from the posterior mean of F_{jk}:

$$\mathrm{PIP}\big(W_{jk}\big) := \Pr\big(W_{jk} \text{ comes from larger effect} | \mathrm{Data}\big) = \Pr\big(F_{jk} = 1 | \mathrm{Data}\big).$$

The sparse factors inferred can then be interpreted, *e.g.* through gene ontology (GO) enrichment analysis of genes loaded on the factors, providing information about the biological effects of perturbations.

While the effects of genetic perturbations are formulated in terms of factors under GSFA, the model does allow us to infer the effects on individual genes. This is similar to the commonly used differential gene expression analysis, where the expression of genes in cells with certain perturbation are compared with those without it. However, when a perturbation affects multiple factors, it can be difficult to synthesize the effects of this perturbation across all affected factors. GSFA provides a way to integrate information over all factors to compute the total effect of a target perturbation on individual genes. The total effect of perturbation m on the expression of gene j, denoted as θ_{mj}, is given by the sum of effects mediated through all K factors:

$$\theta_{mj} = \sum_{k} \beta_{mk} W_{jk}.$$

To sample the posterior distribution of θ_{mj}, we use the posterior samples of β_{mk} and W_{jk} (F_{jk}):

$$\theta_{mj}^{(t)} = \sum_{k=1}^{K} \beta_{mk}^{(t)} W_{jk}^{(t)} F_{jk}^{(t)}, \quad (14)$$

where superscript (t) denotes the t-th posterior sample. The significance of the summarized total effect is evaluated using local false sign rate (LFSR)[117], a metric that is analogous to local false discovery rate (LFDR), but reflects confidence in the sign of effect rather than in the effect being non-zero. It has been shown that LSFR has some benefits over the commonly used FDR approach, and is in fact more conservative than LFDR[117]. The LFSR of the perturbation effect on individual genes, θ_{mj}, can be obtained from its posterior samples by:

$$\mathrm{LFSR}(\theta_{mj}) = \min\{\Pr(\theta_{mj}^{(t)} \geq 0|\mathrm{Data}), \Pr(\theta_{mj}^{(t)} \leq 0|\mathrm{Data})\}. \quad (15)$$

By thresholding LFSR, we can obtain significant DEGs under each perturbation. In practice, the threshold is LFSR < 0.05.

In summary, GSFA produces three main outputs (Fig. 3.1, bottom): the association between genetic perturbations and factors; the weights of genes on factors measured by PIPs; and a list of DEGs of each perturbation at a given LFSR cutoff.

3.3.4 Alternative models for GSFA

Separate estimation of perturbation effects on multiple cell groups

In the cases when one is interested in learning about the effects of perturbations under different cell types or experimental conditions, the base GSFA model can be extended to estimate these effects separately and produce different DEGs for each cell type/condition. Specifically, we modify the current relationship between **Z** and **G** (Equation 11) so that the factors are still inferred using all cells, but the associations between factors and perturbations are estimated separately for different cell groups. For example, assuming 2 groups of cells, group 0 and group 1, the dependency of **Z** on **G** is now modeled as:

$$P(\mathbf{Z}|\mathbf{G}, \boldsymbol{\beta_0}, \boldsymbol{\beta_1}) = \prod_{i0 \in \text{group } 0} N(Z_{i0\cdot}; \boldsymbol{\beta_0} G_{i0\cdot}, \mathbf{I}_K) \cdot \prod_{i1 \in \text{group } 1} N(Z_{i1\cdot}; \boldsymbol{\beta_1} G_{i1\cdot}, \mathbf{I}_K)$$

where $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ are $M \times K$ matrices holding the effect sizes of perturbations on factors within group 0 cells and group 1 cells, respectively. Each effect size matrix is still subject to the same "spike-and-slab" prior:

$$\beta_{0mk} \sim p_{0m} N(0, d_{0m}^2) + (1 - p_{om})\delta_0,$$

$$\beta_{1mk} \sim p_{1m} N(0, d_{1m}^2) + (1 - p_{1m})\delta_0.$$

The distributions of other model parameters remain the same.

Once we have the posterior samples of parameters and latent variables, we can similarly obtain the posterior samples of the total effects of perturbations on individual genes, θ_{mj}'s, within each cell group using Equation (14) and the corresponding β_{mk} for that cell group.

Spike-and-slab prior on gene loadings

GSFA also allows one to use the standard spike-and-slab prior for the gene weights, although we find that it does not work as well as the default normal-mixture prior (Equation 13). The alternative spike-and-slab prior is given by:

$$W_{jk} \sim \pi_k N(0, \sigma_k^2) + (1 - \pi_k)\delta_0. \quad (16)$$

Again, we introduced a latent binary matrix $\mathbf{F}_{P \times K}$ to indicate whether W_{jk}'s are nonzero. Similar to the inference of $\boldsymbol{\beta}$, we can obtain the posterior samples of $\mathbf{F}$ and $\mathbf{W}$ by updating F_{jk} and W_{jk} ($1 \leqslant j \leqslant P$, $1 \leqslant k \leqslant K$) in pairs. We first sample F_{jk} from a Bernoulli distribution following:

$$\frac{P(F_{jk} = 1|\cdot)}{P(F_{jk} = 0|\cdot)} = \sqrt{\frac{\lambda_{jk}}{\sigma_k^2}} \exp\left(\frac{\nu_{jk}^2}{2\lambda_{jk}}\right) \cdot \frac{\pi_k}{1 - \pi_k},$$

where $\nu_{jk} = \lambda_{jk} \sum_{i=1}^N Z_{ik}\left(Y_{ij} - \sum_{h:h \neq k} Z_{ih} W_{jh}\right)/\psi_j$ and $\lambda_{jk} = \left(\sum_{i=1}^N Z_{ik}^2/\psi_j + \sigma_k^{-2}\right)^{-1}$. With F_{jk} sampled, we can obtain posterior samples of W_{jk} with

$$W_{jk}|F_{jk} = 1 \sim N(\nu_{jk}, \lambda_{jk}),$$

$$W_{jk}|F_{jk} = 0 \sim \delta_0.$$

3.4 Application of GSFA on simulated and real datasets

3.4.1 Simulation studies

Simulation setup

We performed extensive simulations to evaluate the performance of GSFA under two settings, with either continuous gene expression levels or discrete gene count data as output.

(1) Normal distribution scenarios

In the first simulation setting, we generated continuous gene expression levels with a normal error distribution, according to the GSFA model.

$$G_{im} \overset{i.i.d}{\sim} \text{Bernoulli}(0.05), \phi_{ik} \overset{i.i.d}{\sim} N(0,1) \rightarrow \mathbf{Z} = \mathbf{G\beta} + \mathbf{\Phi},$$

$$W_{jk} \overset{i.i.d}{\sim} \pi N(0,0.5) + (1-\pi)\delta_0, E_{ij} \overset{i.i.d}{\sim} N(0,1) \rightarrow \mathbf{Y} = \mathbf{ZW}^{\mathrm{T}} + \mathbf{E},$$

Each dataset consists of $N = 4000$ cells, $P = 6000$ genes, $M = 6$ types of perturbations, and $K = 10$ underlying factors. Each perturbation occurs in ~5% of cells, mimicking real multiplex CRISPR screening assays. The proportion of genes with non-zero effects on each factor, π, referred to as factor density below, varies from 5%, 10% to 20% under different simulation scenarios. For simplicity, each perturbation is designed to affect a distinct factor, and the effect size matrix $\boldsymbol{\beta}$ is set to the following form:

$$\boldsymbol{\beta} = \begin{pmatrix} 0.1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.4 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.6 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

The effect sizes vary from 0.1 to 0.6, which means that the perturbations explain about 0.2% to 8% of the total variance of each factor.

(2) Count-based scenarios

The second simulation setting mimics real scRNA-seq UMI data. To sample the read count data, we first assume that each cell has a library size/scaling factor L_i, sampled from a normal distribution with mean 5×10^5. The count of a gene j in cell i is then sampled from a Poisson distribution with its mean determined by the continuous gene expression level Y_{ij} and the scaling factor L_i:

$$L_i \sim N(5 \times 10^5, 10^5) \rightarrow c_{ij} \sim \text{Poisson}\left(L_i \exp\left(1/5 \times 10^5 + Y_{ij}\right)\right).$$

The sampled counts are converted to deviance residuals (3.4.1), then centered and scaled so that each gene has variance 1 before provided as input for GSFA. Other simulation parameters remained the same as in the normal setting.

We generated 300 random datasets under each of the 6 scenarios (normal/count-based and π = 0.05, 0.1, 0.2) for GFSA analysis. For each dataset, Gibbs sampling was performed for 3000 iterations, and posterior means of parameters were computed from the last 1000 iterations, which took 1.1 hours on a modern Linux workstation with Intel Xeon E5-2680 v4 (2.40 GHz) processors.

Comparison of different sparse priors on gene loadings

Lastly, simulated data also allowed us to evaluate the model choice, particularly the prior distribution on gene loadings (**W**) in count-based data. For data simulated under the normal distribution scenario, both the normal-mixture prior (Equation 13) and the spike-and-slab prior (Equation 16) produced factors with similar density as the ground truth (Fig. 3.2a). However, for count-based simulation data, factors inferred under the spike-and-slab prior sometimes resulted in factors much denser than the ground truth, with a small fraction (~10%) of them having a density

level close to 1. In contrast, factors inferred under the normal-mixture prior consistently have sparse gene weights, with the proportion of loaded genes centered around the ground truth (Fig. 3.2b). This justifies our choice of normal-mixture prior as the default prior for read count data.

Evaluation of factor inference and recovery of perturbation effects on factors

We first evaluate the performance of GSFA in the inference of factors. Due to the interchangeability of factors in matrix factorization, we map each of the true factors to the GSFA inferred factor that it is maximally correlated with using the absolute Pearson correlation. The correlations of the true and matching inferred factors are then assessed. Across all scenarios, factors inferred by GSFA are highly correlated with the true underlying factors, with 80-90% of the absolute correlation values above 0.8 (Fig. 3.2c). GSFA also recovers genes with nonzero loading on the factors. Indeed, genes with PIPs above 0.95 are generally true genes, with observed false discovery proportions (FDP) below 0.1 when the true factor density is less than 0.2 (Fig. 3.2d).

Next, we evaluate the performance of GSFA in detecting the effects of perturbations on factors. Across all scenarios, GSFA estimates these effects accurately (Fig. 3.2e). The small downward bias of estimated effects is expected, given the sparse prior we imposed. We further assessed the calibration of PIPs of these effects. At a PIP threshold of 0.95 and a true factor density level below 0.2, the proportion of falsely detected effects is generally below 0.1 (Fig. 3.2f).

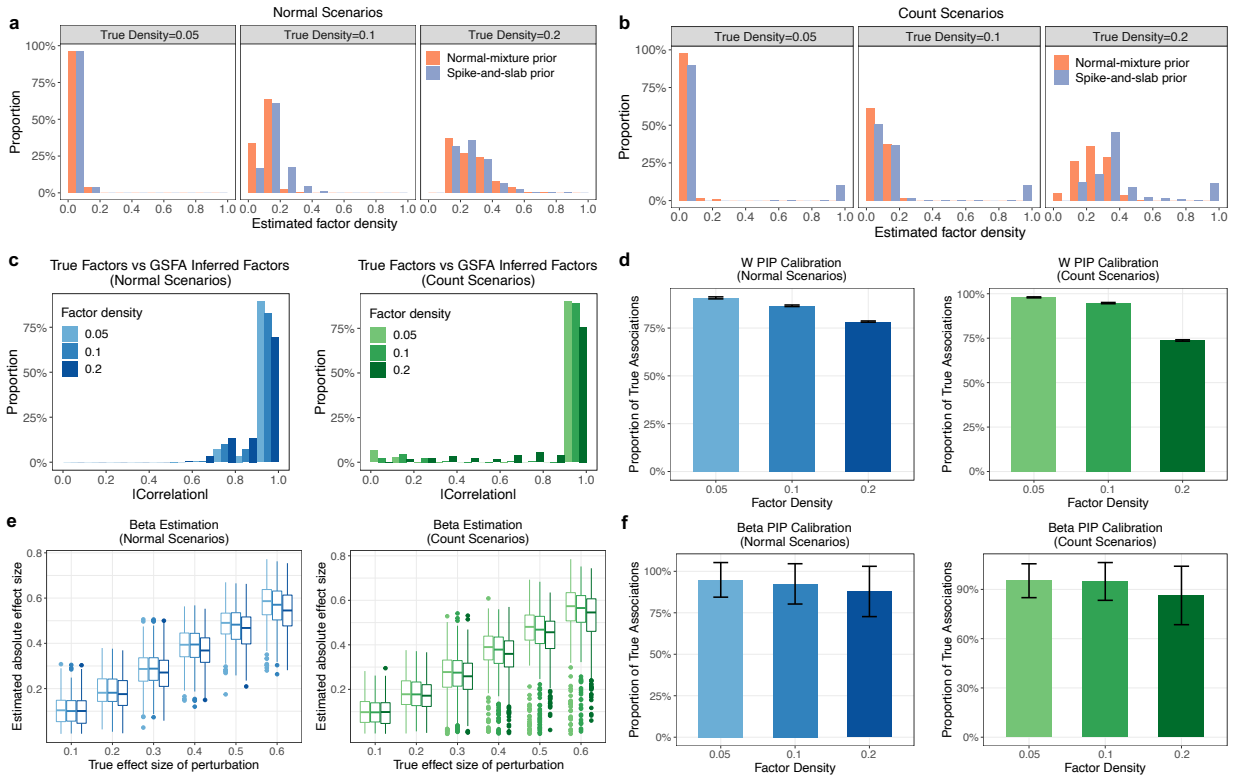

Figure 3.2: GSFA performance on simulated data – factor estimation.
a) Comparison of estimated factor densities using two priors under the normal-based settings. **b**) Comparison of estimated factor densities using two priors under the count-based settings. **c**) Distributions of the absolute correlation values between true factors and the factors inferred by GSFA with true factor density varying from 0.05 to 0.2, under the normal-based or count-based settings. **d**) Left: normal-based settings; right: count-based settings. The proportion of truly associated factor-gene pairs out of all the pairs that have GSFA estimated gene loading PIP > 0.95 in the corresponding factor, computed for each dataset under three levels of true factor density. The range of each error bar is one standard deviation ± mean, computed over the proportion values of 300 datasets. **e**) Left: normal-based settings; right: count-based settings. Box plots of absolute effect sizes from perturbation-factor regression estimated by GSFA. In each plot, different colors represent different values of true factor density varying from 0.05 to 0.2. The center line of a box represents the median; the lower and upper hinges of a box correspond to the first and third quartiles; the upper/lower whisker extends from the hinge to the largest/smallest value no further than 1.5 * inter-quartile range from the hinge. **f**) Left: normal-based settings; right: count-based settings. The proportion of truly associated perturbation-factor pairs out of all the pairs that have GSFA estimated association PIP > 0.95, computed for each dataset under three levels of true factor density. The range of each error bar is one standard deviation ± mean, computed over the proportion values of 300 datasets.

Evaluation of DEG discovery

Finally, we compared the performance of GSFA in DEG detection, *i.e.*, detection of genes affected by perturbations, with commonly used differential expression analysis methods. For a

given perturbation *m*, we define the ground truth of genes affected by it as those with non-zero weights (true $W_{jk} \neq 0$) in the factor that this perturbation is truly associated with (true $\beta_{mk} \neq 0$). These ground truth genes are compared with the DEGs discovered by GSFA under the threshold LFSR < 0.05. For comparison, we applied Welch's t-test[126] to both the normal data and count-normalized deviance residual data. For count data scenarios, we also applied edgeR quasi-likelihood F-test (edgeR-QLF)[73] and MAST[91], a method designed for single-cell analysis. In all these DE tests, cells with each perturbation were compared with all other cells without this perturbation for all genes, FDR was computed following the Benjamini-Hochberg procedure for genes under each perturbation, and significant DEGs were obtained by thresholding FDR < 0.05.

We generated receiver operating characteristic (ROC) curves for each method by varying the LFSR or FDR threshold. GSFA outperformed the other methods in terms of both sensitivity and specificity under all scenarios (Fig. 3.3a, Fig. 3.4, Fig. 3.5). In addition, DEGs detected by GSFA at LFSR < 0.05 have observed false discovery proportions (FDPs) well below 0.05 in most cases, while the observed FDPs among edgeR or t-test DEGs show significant inflation under the count-based scenarios (Fig. 3.3b).

Through these simulations, we have demonstrated that GSFA is a powerful method to identify gene modules, the associations between perturbations and modules, and the specific genes affected by each perturbation.

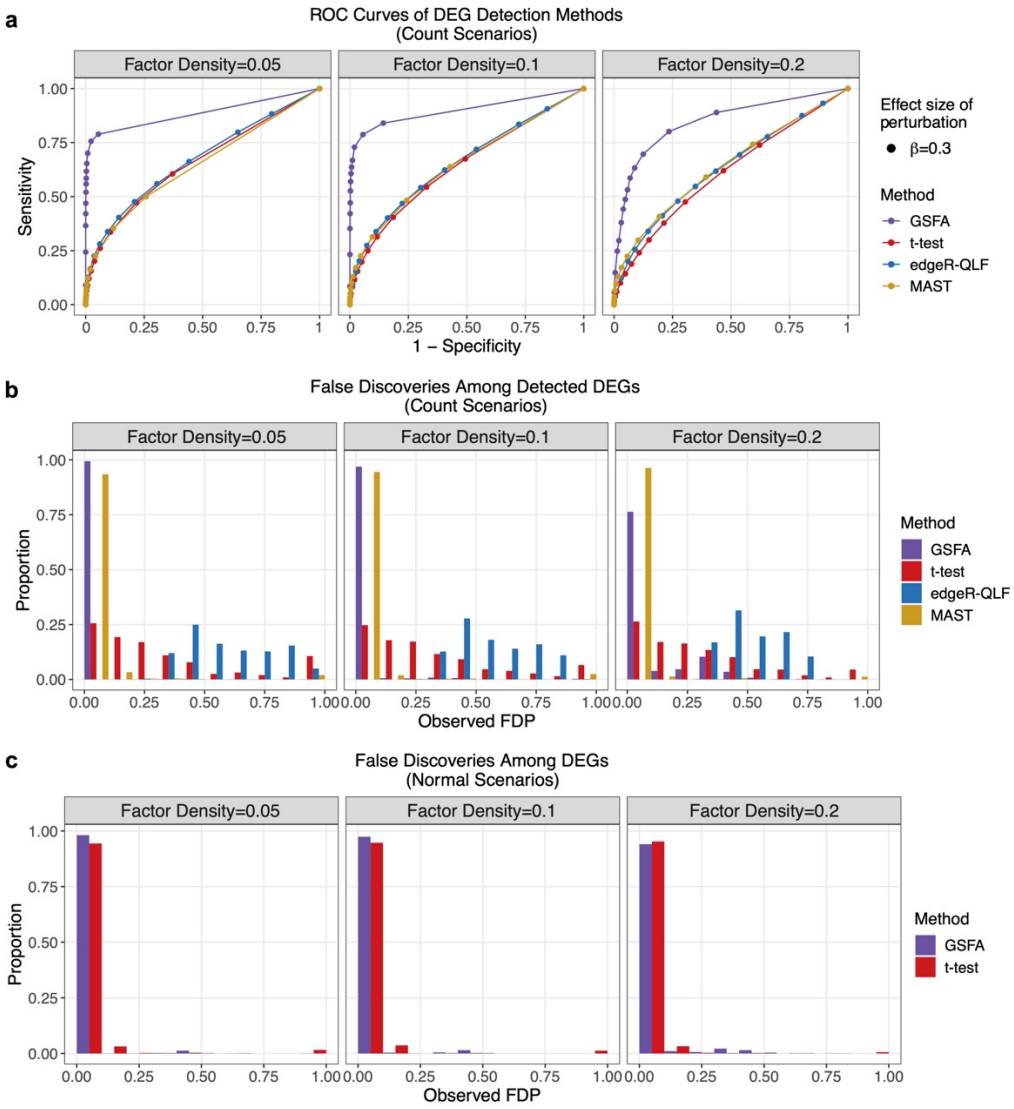

Figure 3.3: GSFA performance on simulated data – DEG discovery.

a) ROC curves of DEG discovery under the count-based setting and 3 different levels of true factor density; 4 colors correspond to 4 DEG detection methods. Results shown are of perturbations with a true association effect of 0.3 on factors. Each was an average over 300 datasets generated under the corresponding setting. See Fig. S2 and S3 for results under other settings. **b**) Distributions of observed proportion of false discoveries (FDP) among significant DEGs detected by GSFA (LFSR < 0.05) and other methods (FDR < 0.05) per dataset under the count-based setting and various true factor densities. 4 colors correspond to 4 DEG detection methods. **c**) Same as in b) but under the normal settings; 2 colors correspond to 2 DEG detection methods.

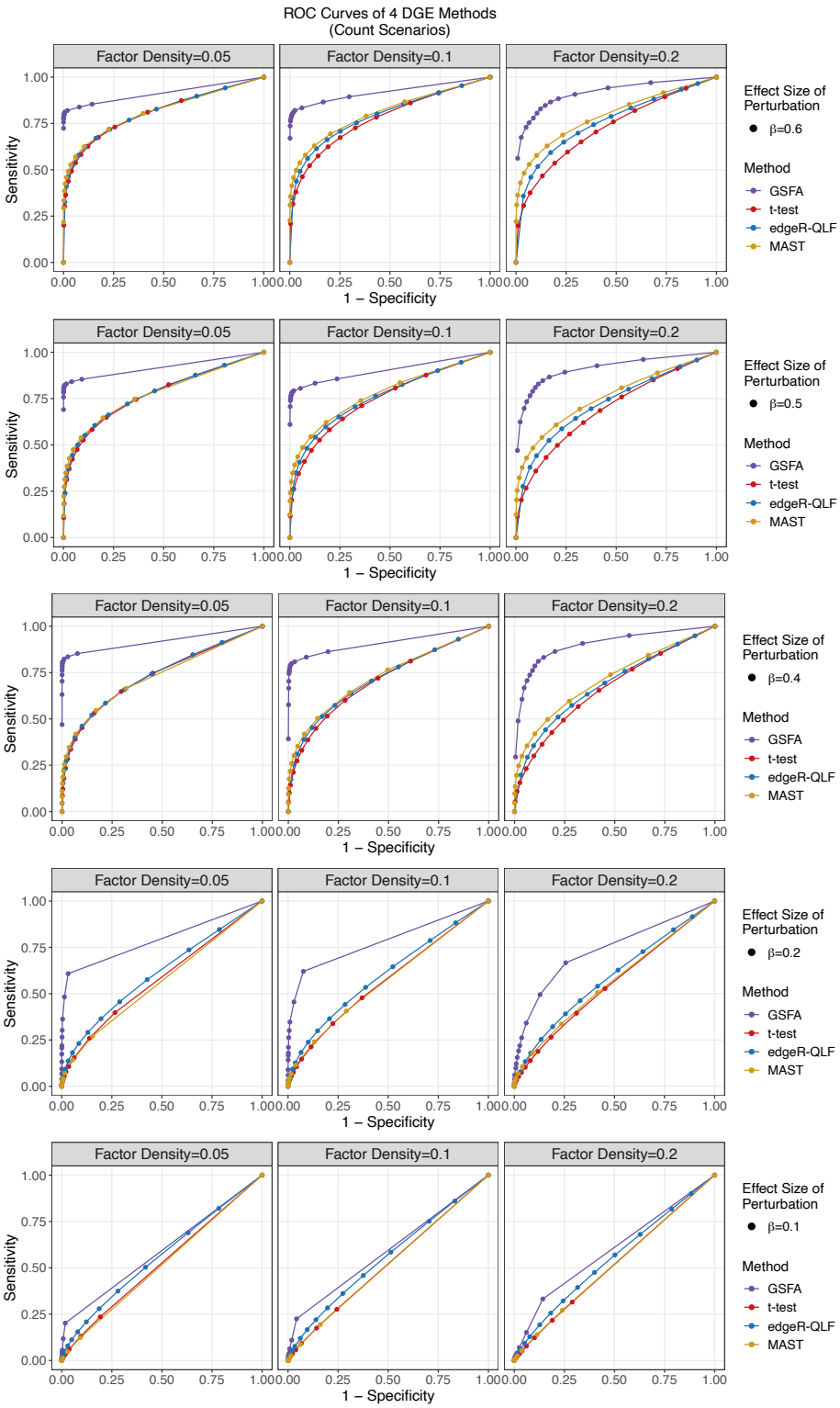

Figure 3.4: ROC curves of DEG discovery across methods on count-based simulated data.
Results are across 3 different levels of true factor density, and 4 different values of true perturbation effects; 4 colors correspond to 4 DEG detection methods.

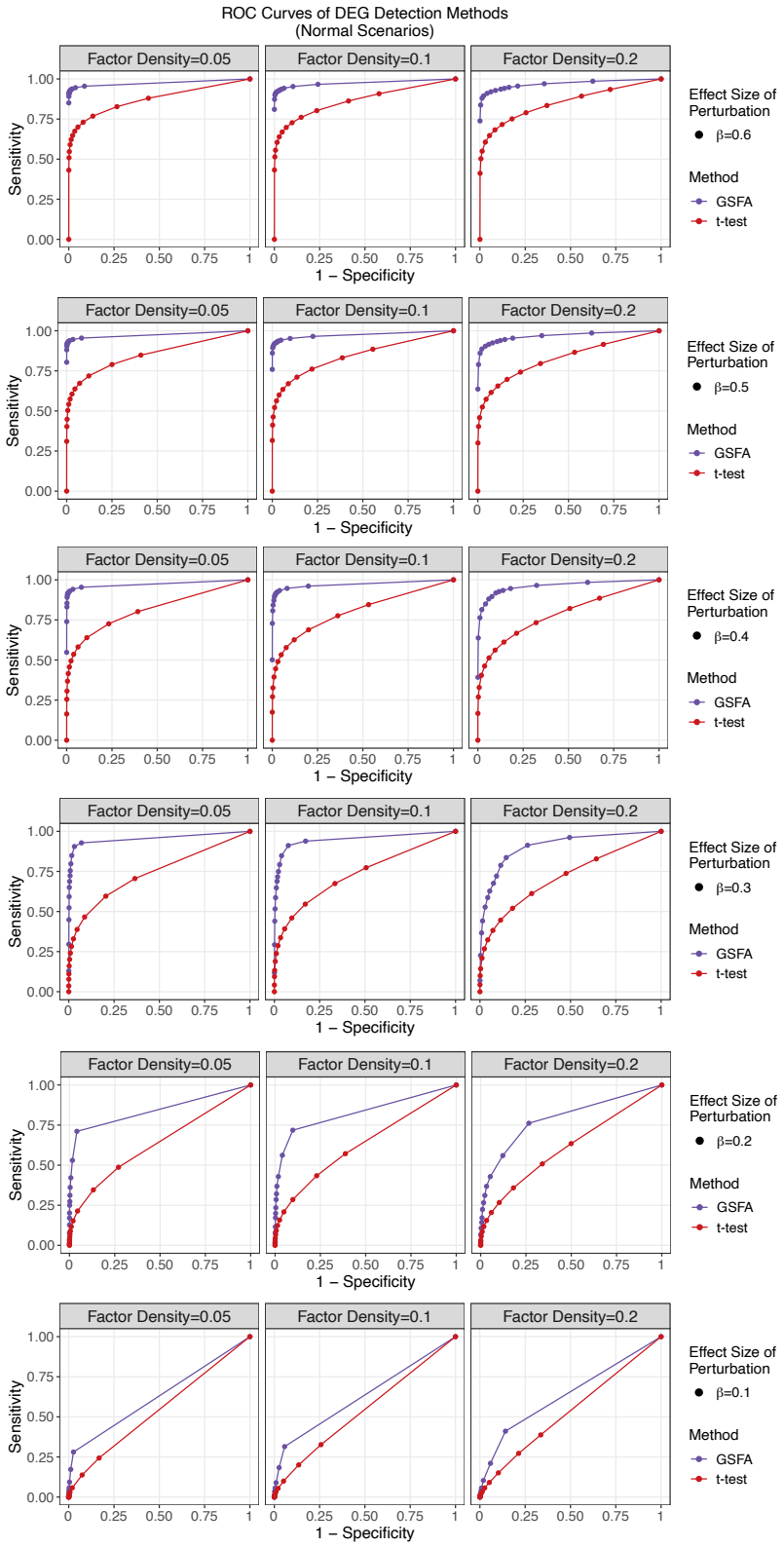

Figure 3.5: ROC curves of DEG discovery across methods on normal scenario simulated data. Results are across 3 different levels of true factor density, and 5 different values of true perturbation effects; 2 colors correspond to 2 DEG detection methods.

3.4.3 GSFA application on human CD8$^+$ T cell CROP-seq data

The first real dataset we applied GSFA to is a CROP-seq dataset of perturbed primary human CD8$^+$ T cells with or without T cell receptor (TCR) stimulation[95]. In the original study, a genome-wide pooled CRISPR screen was conducted to uncover regulators of T cell proliferation upon TCR stimulation. Next, a CROP-seq experiment was carried out targeting a total of 20 genes, including 12 genes that were top hits in the genome-wide screen, and 8 known immune checkpoint genes. The perturbed cells were either TCR-stimulated or not before sequencing, with the purpose of understanding the impact of these genetic perturbations on the transcriptional states of cells in both resting state and stimulated state. The original study applied a clustering approach to characterize the effects of each perturbation. Although perturbations of some genes were found to be correlated with clusters characterized by T cell activation or resting states, many other genes were not associated with any cluster. Moreover, the study lacked systematic differential expression analysis to reveal specific genes affected by perturbations.

To fill in these gaps, we applied GSFA to this CROP-seq dataset, allowing perturbations to have different effects on factors in stimulated and unstimulated cells. With a total of 20 factors pre-specified in GSFA, we obtained 24 associations (PIP > 0.95) between perturbations and factors in stimulated cells that involved 9 gRNA-targeted genes (Fig. 3.6a). Among these genes, the effects of *ARID1A*, *DGKA*, *SOCS1*, and *TCEB2* were undetected by clustering analysis in the original study (Fig. 3.6b). In unstimulated cells, only three pairs of associations were detected at PIP > 0.95 (Fig. 3.6c), which is unsurprising given the role of these targeted genes in regulating T cell responses. Nevertheless, it is interesting that some perturbations (*e.g. TCEB2*, *RASA2*) have effects in unstimulated cells (Fig. 3.6c), suggesting that these genes affect resting state transcriptomes. We also confirmed, with permutation analysis, that the full GSFA results,

including the inferred perturbation effects and gene loading, were calibrated (Methods, Fig. 3.8a-d). Altogether, these results highlight the power of GSFA to detect the broad effects of target genes on the latent factors.

To characterize these latent factors, particularly those associated with perturbations, we inspected the weights of some canonical marker genes (Table 3.1) and performed gene ontology (GO) enrichment analysis of genes loaded on the factors. As an example, factors 2 and 9 have negative weights for cell proliferation markers *MKI67*, *TOP2A*, and *CENPF* (Fig. 3.6d), and are enriched for GO terms related to cell cycle and division (Fig. 3.6e). Factors 4 and 12 are associated with markers of T cell activation and/or resting states (Fig. 3.6d) and are enriched for GO terms related to immune responses (Fig. 3.6e). Together, these results show that the latent factors discovered by GSFA represent cellular processes.

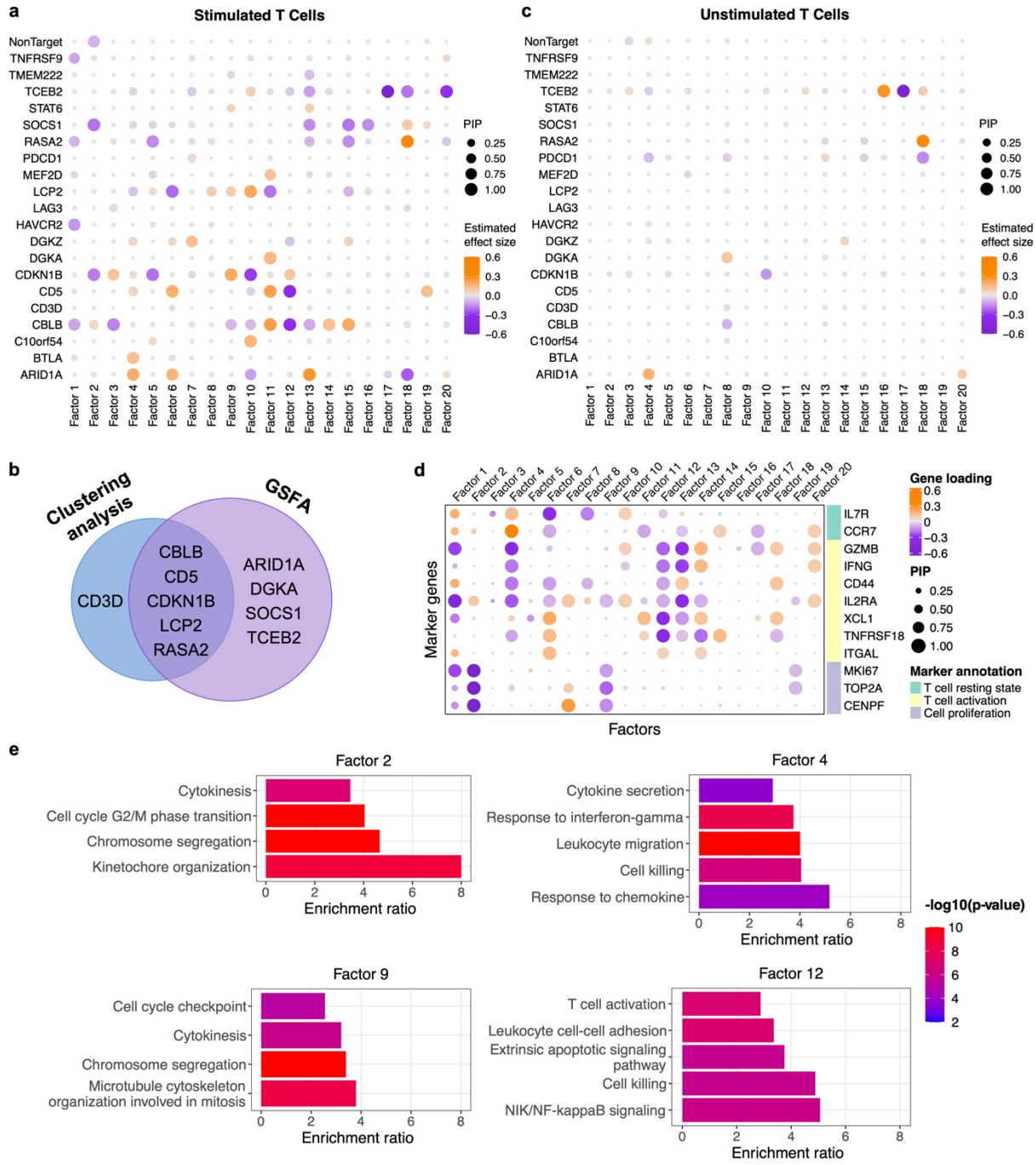

Figure 3.6: GSFA results of inferred factors from analysis of CROP-seq data of primary CD8$^+$ T cells. **a**) Estimated effects of gene perturbations on all factors inferred by GSFA within stimulated T cells. The size of a dot represents the PIP of association; the color represents the effect size. **b**) Venn diagram of targets identified using the original clustering-based method vs. GSFA in stimulated T cells. **c**) Similar to a) but estimated within unstimulated T cells. **d**) Loading of selected marker genes on inferred factors. The size of a dot represents the gene PIP in a factor and the color represents the gene weight (magnitude of contribution) in a factor. **e**) The fold of enrichment for selected GO "biological process" gene sets significantly enriched (q-value < 0.05) in factor 2, 4, 9, and 12. Hypergeometric test was used, where genes with PIP > 0.95 in the factor were compared against a background of all genes used in GSFA.

Gene	Protein (Aliases)	Annotation	References
IL7R	Interleukin-7 receptor (CD127)	T cell resting state	PMID: 15308108
CCR7	CC chemokine receptor 7	T cell resting state	PMID: 11145663
GZMB	Granzyme B	T cell activation	PMID: 12360212, PMID: 22084442
IFNG	Interferon gamma	T cell activation	PMID: 11145690
CD44		T cell activation	PMID: 12526810
IL2RA	Interleukin-2 receptor	T cell activation	PMID: 18417224
XCL1	X-C motif chemokine ligand 1	T cell activation	PMID: 19913446
TNFRSF18	Glucocorticoid-induced TNFR-related protein (GITR)	T cell activation	PMID: 21076066
ITGAL	Integrin subunit alpha L (LFA-1)	T cell activation	PMID: 29774029
MKI67	Marker of proliferation Ki-67	Cell proliferation	PMID: 29322240
TOP2A	DNA topoisomerase II alpha	Cell proliferation	PMID: 15980158
CENPF	Centromere protein F	Cell proliferation	PMID: 16565862

Table 3.1: Details of selected T cell marker genes

We observed that some perturbations are associated with multiple factors with similar functions, *e.g.*, targeting of *CDKN1B* is negatively associated with factor 2 and positively associated with factor 9, and both are enriched for cell cycle GO terms (Fig. 3.6a,e). In this case, it would be hard to consolidate the effects of a target on multiple overlapping factors using a typical two-step approach where factor analysis is followed by target-factor association analysis. GSFA, on the other hand, provides a way to summarize all these factor-mediated effects and estimate a total perturbation effect on each individual gene evaluated by LFSR, demonstrating its unique advantage in DEG detection over traditional factor analysis models.

Besides GSFA's LFSR-based analysis to identify specific downstream genes affected by the perturbations, we also ran several other differential expression analysis methods for comparison, including scMAGeCK-LR[27], a method tailored for single-cell CRISPR screening data, MAST[91], DESeq2[90], and edgeR-QLF[73]. scMAGeCK-LR and MAST had calibrated false positive rates in the permuted data, while DEseq2 showed modest inflation and edgeR-QLF severe

inflation (Methods, Fig. 3.8e-h). Therefore, we excluded edgeR-QLF from further analysis. In stimulated T cells, GSFA detected 148 to 710 DEGs at LFSR < 0.05 for 9 gene targets, four of which (*ARID1A*, *DGKA*, *SOCS1*, and *TCEB2*) were poorly characterized by clustering analysis in the original study[95]. Compared with other methods, GSFA consistently detected the most DEGs across these 9 targets, sometimes with 10 times or more DEGs (Fig. 3.7a). Additionally, the DEGs of all 9 target genes detected by GSFA are enriched for a larger number of GO terms (Fig. 3.7b), many of which are relevant to T cell responses (Fig. 3.7c). DEGs detected by other methods, in contrast, have almost no GO enrichment except for *TCEB2*. These results thus show that GSFA has much higher sensitivity in detecting DEGs than existing methods.

With a large number of DEGs detected for the 9 target genes, we next focused on their effects on specific marker genes of T cell activation to better understand the genes' functions. GSFA revealed a number of effects of the target genes on the markers (Fig. 3.7d), many of which were missed by other methods (Fig. 3.7f-h). The estimated effects of these genes on the chosen markers largely agree with the reported roles of these genes[95]. For instance, targeting of *CD5*, *CBLB* and *RASA2* have mostly positive effects on markers of activated T cells, and negative or no effects on markers of resting T cells (Fig. 3.7d), consistent with the functions of these genes as negative regulators of T cell stimulation[95]. Targeting of *CDKN1B* has strong positive effects on cell proliferation markers (Fig. 3.7d), consistent with its function in the cell cycle[127] and its role as a negative regulator of T cell proliferation[95].

Our analysis also revealed molecular effects of the four novel genes, *ARID1A*, *DGKA*, *SOCS1*, and *TCEB2*, whose effects were poorly characterized in the earlier study (Fig. 3.6b). The effects of perturbations of *TCEB2* and *DGKA* on T cell markers are similar to those of other negative regulators of T cell responses, such as *CD5*, and were found to be negative regulators of T cell

proliferation in the genome-wide screen[95]. Indeed, *TCEB2* encodes Elongin-B, which binds to SOCS1 and forms a complex that attenuates Jak/STAT signaling, negatively regulating cytokine signaling[128]. In addition, as part of the transcription factor B (SIII) complex, *TCEB2* is essential in activating elongation by RNA polymerase II[129]. Consequently, knockout of *TCEB2* has broad transcriptomic impacts beyond immune response processes and independent of TCR stimulation, as observed in the enrichment of RNA metabolic and catabolic processes in its DEGs in both stimulated T cells and unstimulated T cells.

Targeting *SOCS1* has a strong effect on cell proliferation markers (Fig. 3.7d). Accordingly, several genes of the SOCS (suppressor of cytokine signaling) family have been reported to inhibit cell cycle progression[130]. Targeting of *ARID1A*, a chromatin remodeler and potential tumor suppressor[131]–[133], has strong negative effects on the effector markers (Fig. 3.7d), suggesting its role as a positive regulator of T cell activation. Indeed, *ARID1A* mutations occur in many human cancer types, and result in limited chromatin accessibility and down-regulation of interferon-responsive genes, leading to poor tumor immunity[134].

Targeting *LCP2* has strong negative effects on activation markers such as *XCL1* and *TNFRSF18*, and a positive effect on the resting state marker *IL7R*, consistent with the findings of *LCP2* being a positive regulator in the genome-wide screen[95], but the positive effects observed on effector markers such as *GZMB* are unexpected (Fig. 3.7d).

Collectively, GSFA revealed detailed transcriptional effects of 9 genetic perturbations in stimulated $CD8^+$ T cells, including four genes largely missed by clustering or differential expression analysis with other tools. The functions of these four genes suggested by GSFA analysis are largely consistent with their known biological roles. We constructed a regulatory network to summarize our major findings of the functions of all nine target genes (Fig. 3.7e). Our results

highlight the power of GSFA in revealing the detailed molecular effects of genetic perturbations in single-cell CRISPR screens.

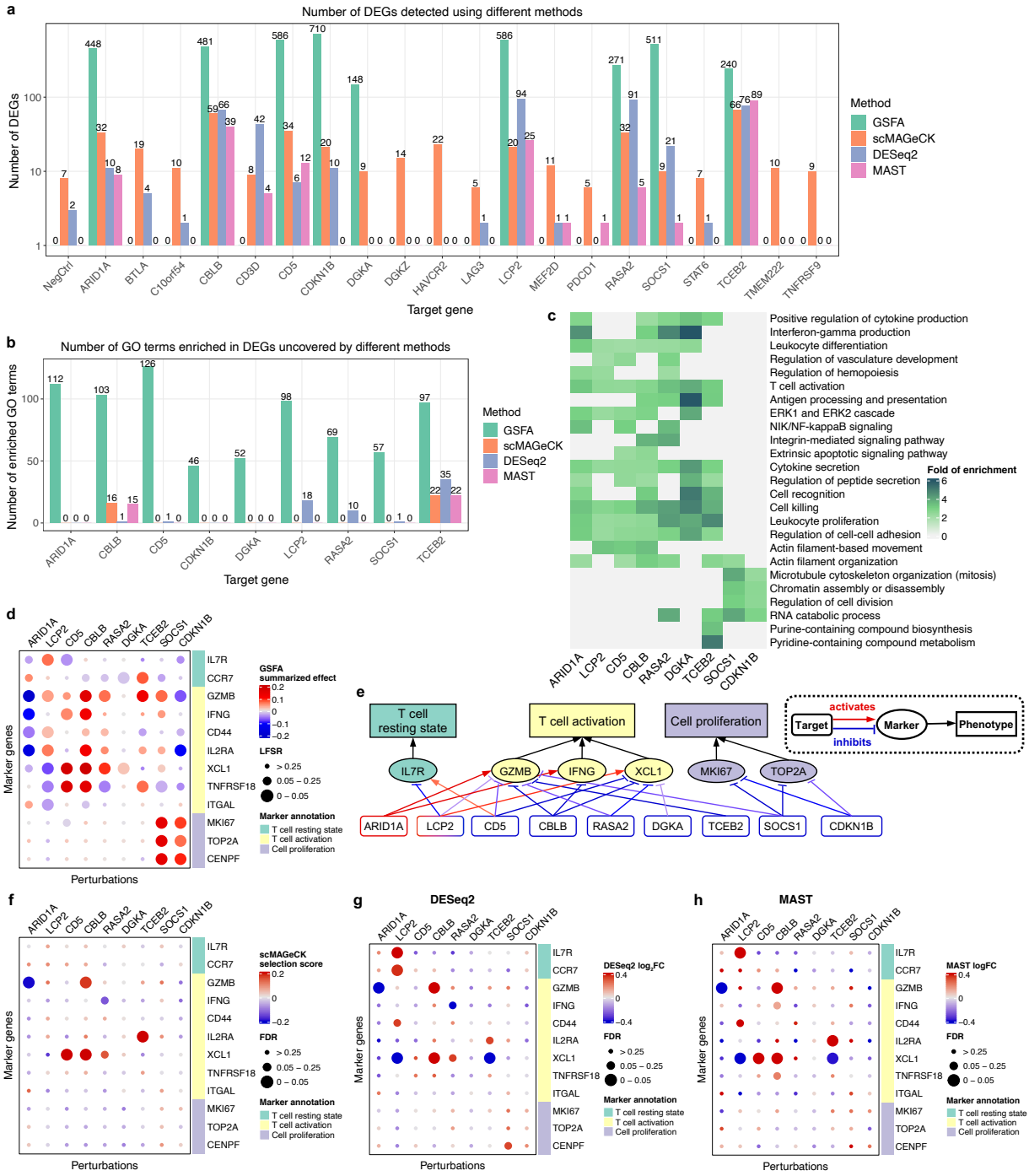

Figure 3.7: GSFA results of the effects of genetic perturbations on gene expression in CD8$^+$ T cell CROP-seq data. The results were based on stimulated CD8$^+$ T cells.

a) Number of DEGs detected under all perturbations using 4 different methods. The y axis is log scaled and bar height corresponds to count+1 (as the number of DEGs could be 0);

(Figure 3.7 continued)
the exact number of DEGs are labeled on top of the bars. The detection threshold for DEGs is LFSR < 0.05 for GSFA, and FDR < 0.05 for all other methods. **b**) Number of GO Slim "biological process" terms enriched in DEGs detected by different methods. **c**) Heatmap of selected GO "biological process" terms and their folds of enrichment in DEGs (LFSR < 0.05) detected by GSFA under different perturbations. **d**) GSFA estimated effects of perturbations on marker genes in stimulated T cells. Sizes of the dots represent LFSR bins; colors of the dots represent the summarized effect sizes. **e**) A target-marker-phenotype regulatory network summarizing GSFA results. Significant (LFSR < 0.05) regulatory relationships between target and marker genes are represented by colored arrows, with red sharp arrows indicating positive regulation by the target genes, and blue blunt arrows indicating negative regulation. The darkness of color represents the relative magnitude of effect. Note the effect directions here are the opposite of the perturbation effects. **f**), **g**), **h**) Estimated effects of perturbations on marker genes in stimulated T cells with scMAGeCK (f), DESeq2 (g), and MAST (h). Sizes of the dots represent FDR bins; colors of the dots represent scMAGeCK selection scores, DESeq2 $\log_2$ fold change estimates, or MAST log fold change estimates, respectively.

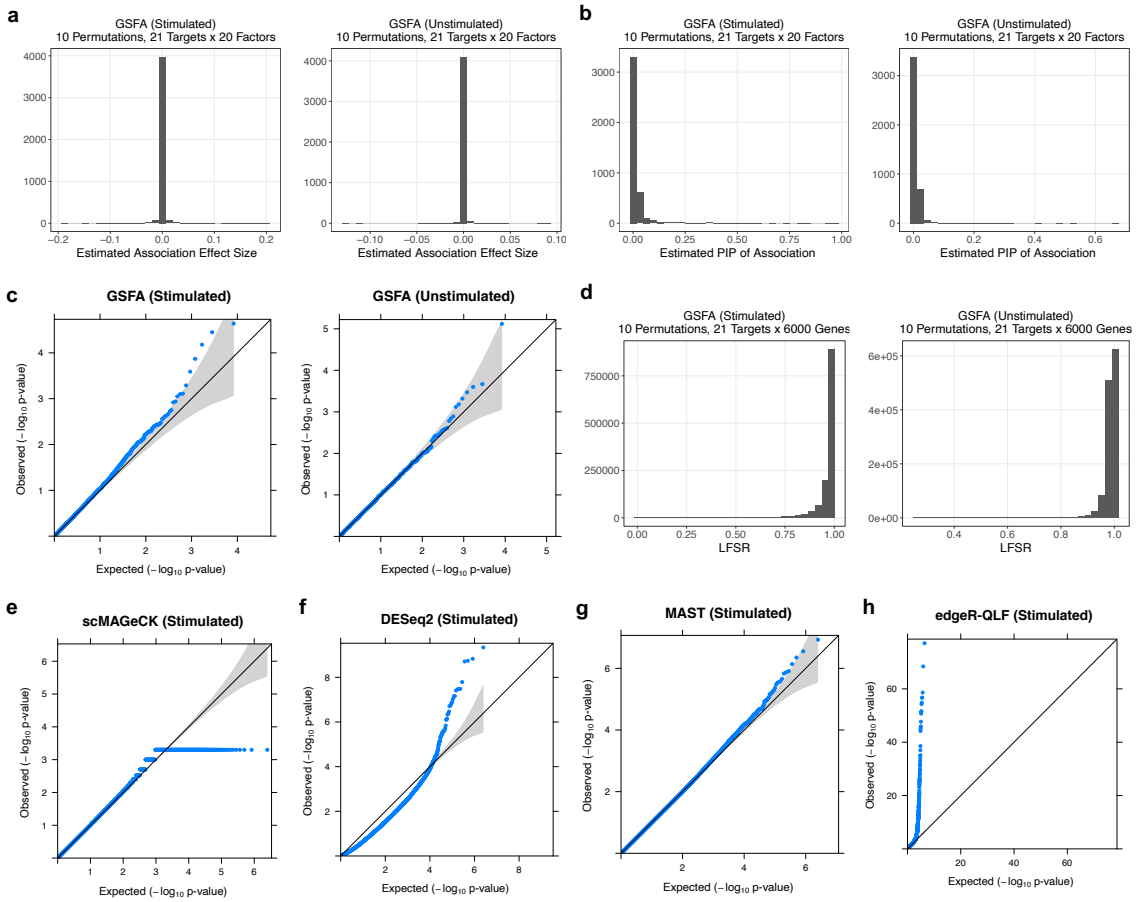

Figure 3.8: Permutation results of DEG detection methods on CD8$^+$ T cell CROP-seq data. Results from 10 randomly permuted datasets are pooled together. **a)** GSFA effect sizes of perturbations on factors, estimated within stimulated cells and unstimulated cells, respectively. **b)** GSFA PIPs of associations between factors and perturbations, estimated within stimulated cells and unstimulated cells, respectively. **c)** Q-Q plot of p-values obtained from linear regression between GSFA estimated factors and perturbations

(Figure 3.8 continued)
within stimulated cells and unstimulated cells, respectively. **d)** GSFA LFSRs of genes under all perturbations, estimated within stimulated cells and unstimulated cells. **e)** Empirical p-values of differential expression estimated by scMAGeCK-LR within stimulated cells; exact zeros were replaced with 5e-4 for visualization in the Q-Q plot. **f)**, **g)**, **h)** Differential expression p-values estimated within stimulated cells by DESeq2 (f), MAST (g), and edgeR-QLF test (h).

3.4.4 GSFA application on LUHMES CROP-seq data

The second real dataset we applied GSFA to is a CROP-seq dataset targeting 14 neurodevelopmental genes, including 13 autism risk genes, in LUHMES human neural progenitor cells[135]. After CRISPR targeting, the cells were differentiated into postmitotic neurons and sequenced, with the purpose of understanding the functions of these risk genes in a neurodevelopmental context. In the original study, cells were projected onto a pseudotime trajectory, which approximates the progression of neuronal differentiation, and the perturbation conditions were associated with the pseudotime of cells. The study concluded that the repression of *ARID1B*, *ASH1L*, *CHD2*, and *DYRK1A* impeded neuronal differentiation, while *PTEN* and *CHD8* repression accelerated neuronal differentiation. However, it provided limited information on the molecular processes affected by the target genes other than pseudotime, and its differential expression analysis did not take measures to control the false discovery rate.

After applying GSFA to this dataset, we found significant effects (PIP > 0.95) of 7 target genes, including *ADNP*, *ARID1B*, *ASH1L*, *CHD2*, *DYRK1A*, *PTEN*, and *SETD5*, on at least one out of 20 latent factors (Fig. 3.9a). Among the 7 genes, the transcriptomic effects of *ADNP* and *SETD5* were missed in the original pseudotime-based analysis (Fig. 3.9b). We also confirmed, that the GSFA results were calibrated as it did not produce false positive findings in permutations (Methods, Fig. 3.11a-d). We characterized these factors by inspecting the weights of neuronal markers (Table 3.2) and GO enrichment analysis of genes loaded on them. In factor 4, for example, the markers of mature neurons such as *MAP2* and *NEFL* have negative weights, while negative

regulators of neuron projection such as *ITM2C* have positive weights (Fig. 3.9c), suggesting that factor 4 is negatively associated with neuronal maturation. Indeed, factor 4 is significantly enriched for gene sets involved in neuronal development (Fig. 3.9d). Factors 9 and 16, similarly, show loadings of neuronal markers and are enriched for relevant GO terms (Fig. 3.9c,d). These results suggest that GSFA was able to relate genetic perturbations with biologically meaningful latent factors.

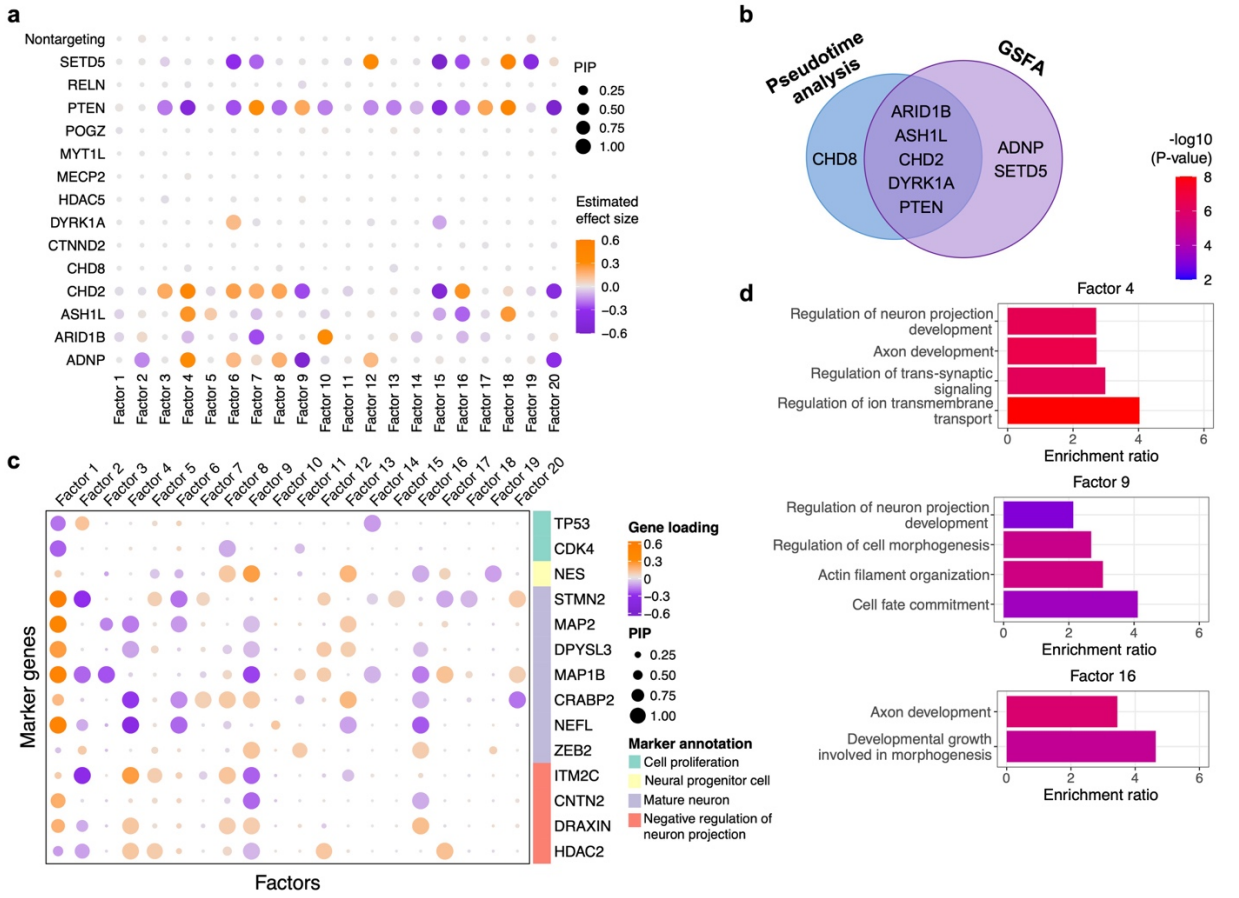

Figure 3.9: GSFA results of inferred factors from analysis of LUHMES CROP-seq data.
a) Estimated effects of gene perturbations on all factors inferred by GSFA. The size of a dot represents the PIP of association; the color represents the effect size. **b**) Venn diagram of targets identified from the original pseudotime association analysis vs. from GSFA. **c**) Loading of neuronal marker genes on factors. The size of a dot represents the gene PIP in a factor and the color represents the gene weight (magnitude of contribution) in a factor. **d**) The fold of enrichment for selected GO "biological process" terms enriched in factor 4, 9, and 16 (q-value < 0.05). Hypergeometric test was used, where genes with PIP > 0.95 in the factor were compared against a background of all genes used in GSFA.

Gene	Protein (Aliases)	Annotation	References
TP53	Tumor protein p53	Cell proliferation	PMID: 18948956
CDK4	Cyclin dependent kinase 4	Cell proliferation	PMID: 19733543
NES	Nestin	Neural progenitor cell	PMID: 29541793
STMN2	Stathmin-2	Mature neuron	PMID: 14598370
MAP2	Microtubule associated protein 2	Mature neuron	PMID: 10704996
DPYSL3	Dihydropyrimidinase like 3	Mature neuron	GO:0010976
MAP1B	Microtubule associated protein 1B	Mature neuron	GO:0010976
CRABP2	Cellular retinoic acid binding protein 2	Mature neuron	GO:0010976
NEFL	Neurofilament Light Chain	Mature neuron	GO:0010976
ZEB2	Zinc finger E-box binding homeobox 2	Mature neuron	GO:0010976
ITM2C	Integral membrane protein 2C	Negative regulation of neuron projection	GO:0010977
CNTN2	Contactin-2	Negative regulation of neuron projection	GO:0010975
DRAXIN	Dorsal inhibitory axon guidance protein	Negative regulation of neuron projection	GO:0010977, PMID: 24832731
HDAC2	Histone deacetylase 2	Negative regulation of neuron projection	GO:0010977
GO:0010975: gene ontology “regulation of neuron projection development” process			
GO:0010976: gene ontology “positive regulation of neuron projection development” process			
GO:0010977: gene ontology “negative regulation of neuron projection development” process			

Table 3.2: Details of neuronal marker genes

We next identified the individual genes affected by the perturbations. GSFA detected DEGs at LFSR < 0.05 for the same 7 gene targets (Fig. 3.10a). Compared with other differential expression analysis methods, GSFA detected the most DEGs for 5 out of 7 gene targets (Fig. 3.10a). Furthermore, DEGs detected by GSFA are enriched for the most GO terms across almost all targets (Fig. 3.10b), many of which are related to neuronal development or neural signaling (Fig. 3.10c).

To better understand the functions of these 7 target genes, we examined their effects on marker genes for neuron maturation and differentiation. GSFA uncovered perturbation effects on a number of neuronal marker genes across all targets except *ARID1B* (Fig. 3.10d), while other methods

detected fewer differentially expressed markers (Fig. 3.10e-g). GSFA-estimated effects of the target genes largely validated the known functions of these genes. Targeting of *ASH1L*, *CHD2*, and *DYRK1A* has mostly negative effects on mature neuronal markers, and positive effects on negative regulators of neuron projection (Fig. 3.10d), indicating delayed neuron maturation by the repression of these genes. Knockdown of *PTEN* showed opposite effects by GSFA, suggesting its opposite role on neuronal differentiation. These results agree with the experimental findings of the effects of these perturbations on neuronal maturation phenotypes[135].

Two genes, *ADNP* and *SETD5*, were missed in the pseudotime-based analysis in the original study (Fig. 3.9b). The estimated effects of these genes on neuronal markers by GSFA suggested that repression of *ADNP* would lead to delayed neuronal differentiation, whereas *SETD5* repression would have the opposite effect (Fig. 3.10d). These predictions are consistent with the experimental finding of *ADNP* in the original study[135], and with the finding that *SETD5* knockdown increases the proliferation of cortical progenitor cells and neural stem cells[136].

In conclusion, GSFA allowed us to identify and characterize the transcriptional effects of 7 ASD risk genes, including *ADNP* and *SETD5*, whose effects were largely missed in the original study. Detailed characterization of these genes by GSFA using neuronal markers provides functional insight that is largely consistent with their known roles in literature. While GSFA missed the effect of *CHD8* (Fig. 3.9b), we noticed that all the existing DEG methods largely missed its effect as well (Fig. 3.10a). We summarized the inferred target effects by GSFA on selected marker genes and affected cellular processes in a gene regulatory network (Fig. 3.10h). Together, these results demonstrate that GSFA is an effective computational tool for single-cell CRISPR screening data in detecting the transcriptional effects and elucidating the molecular mechanisms of genetic perturbations.

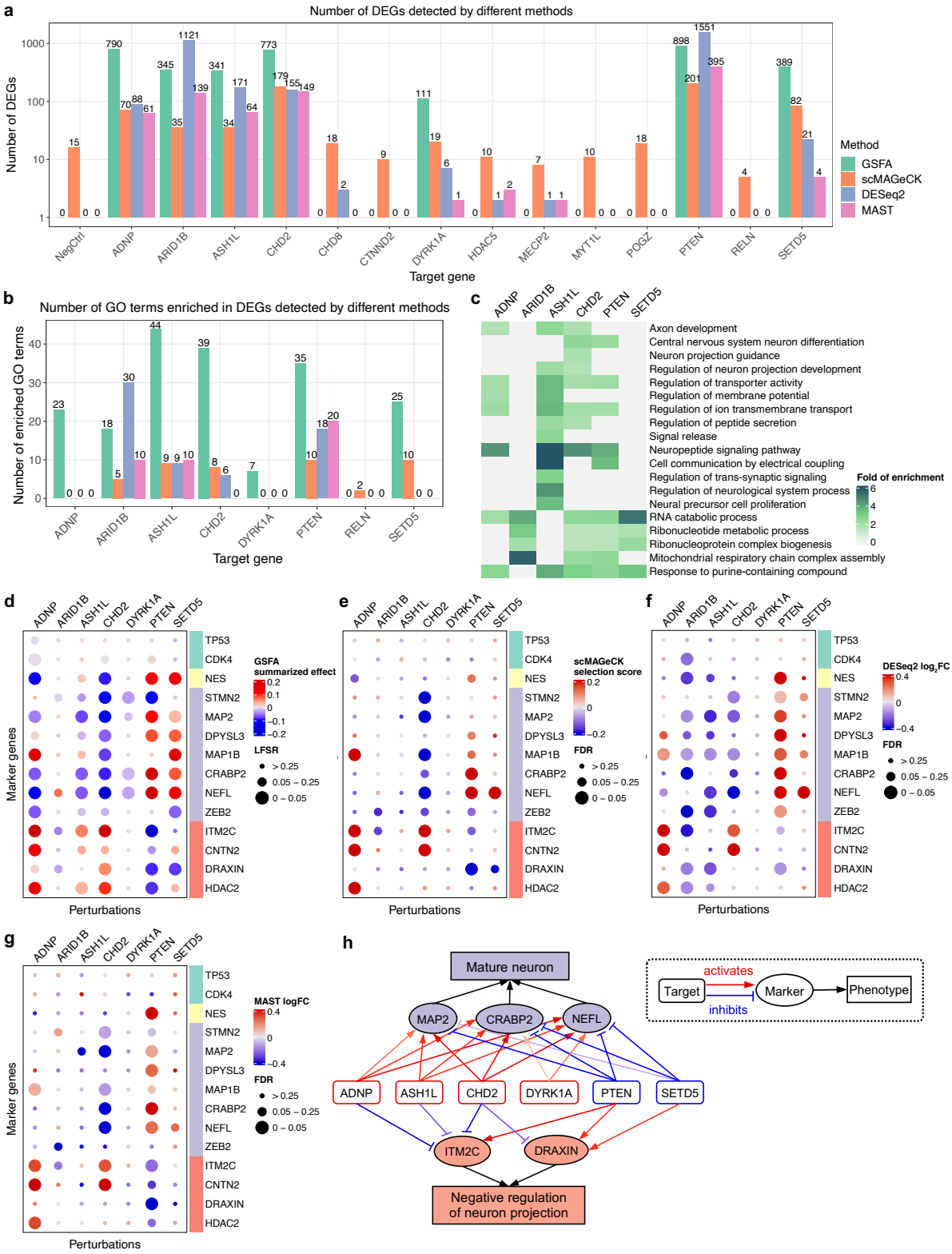

Figure 3.10: GSFA results of the effects of genetic perturbations on gene expression in LUHMES CROP-seq data.

a) Number of DEGs detected under all perturbations using 4 different methods. The y axis is log scaled and bar height corresponds to count+1 (as the number of DEGs could be 0);

(Figure 3.10 continued)
the exact number of DEGs are labeled on top of the bars. The detection threshold for DEGs is LFSR < 0.05 for GSFA, and FDR < 0.05 for all other methods. **b**) Number of GO Slim "biological process" terms enriched in DEGs detected by different methods. **c**) Heatmap of selected GO "biological process" terms and their folds of enrichment in DEGs (LFSR < 0.05) detected by GSFA under different perturbations. **d**) GSFA estimated effects of perturbations on marker genes. Sizes of the dots represent LFSR bins; colors of the dots represent the summarized effect sizes. **e**), **f**), **g**) Estimated effects of perturbations on marker genes with scMAGeCK (e), DESeq2 (f), and MAST (g). Sizes of the dots represent FDR bins; colors of the dots represent scMAGeCK selection scores, DESeq2 $\log_2$ fold change estimates, or MAST log fold change estimates, respectively. **h**) A target-marker-phenotype regulatory network summarizing GSFA results. Significant (LFSR < 0.05) regulatory relationships between target and marker genes are represented by colored arrows, with red sharp arrows indicating positive regulation of marker genes by the target genes, and blue blunt arrows indicating negative regulation. The darkness of color represents the relative magnitude of effect. Note that the direction of regulation is the opposite of the perturbation effect.

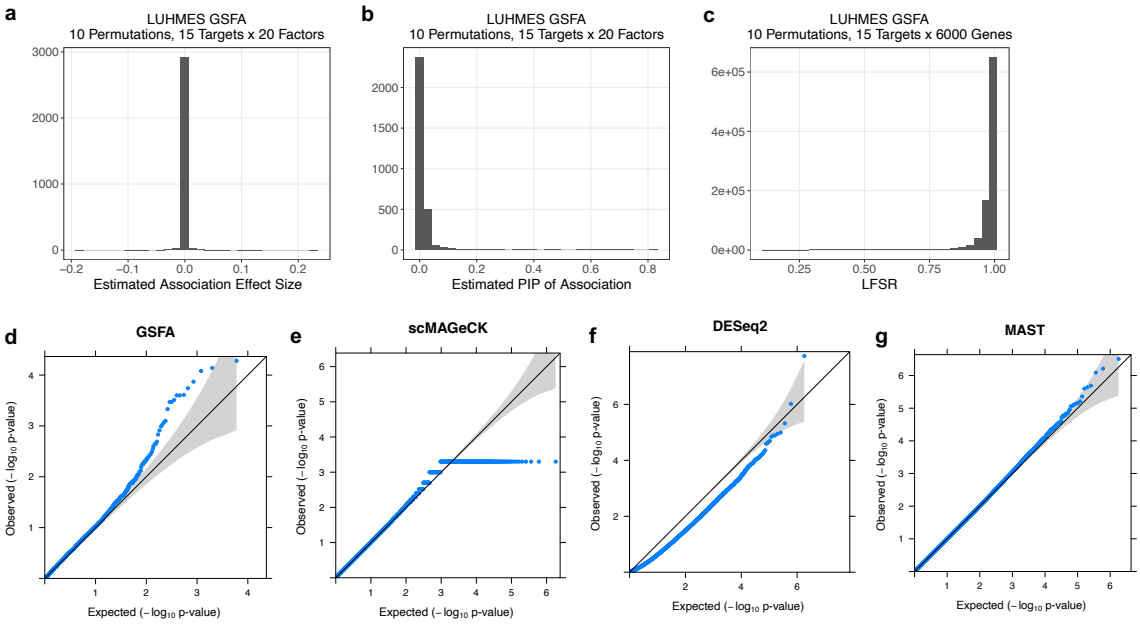

Figure 3.11: Permutation results of DEG detection methods on LUHMES CROP-seq dataset. Results from 10 randomly permuted datasets are pooled together. **a)** GSFA effect sizes of perturbations on factors. **b)** GSFA estimated PIPs of associations between factors and perturbations. **c)** GSFA estimated LFSRs of genes under all perturbations. **d)** Q-Q plot of p-values obtained from linear regression between GSFA estimated factors and perturbations. **e)** Empirical p-values of differential expression estimated by scMAGeCK-LR; exact zeros were replaced with 5e-4 for visualization in the Q-Q plot. **f**), **g)** Differential expression p-values estimated by DESeq2 (f) and MAST (g).

3.4.5 GSFA application on GTEx data

We also explored the application of GSFA on bulk RNA-seq data with paired perturbation information. We obtained from the GTEx (Genotype Tissue Expression) project[137] whole blood

gene expression data and paired individual genotypes at selected SNPs, which serve as naturally occurring genetic perturbations. With the intention to include SNPs with potentially large effect sizes on gene expression modules, we selected a set of 15 independent risk SNPs for GWAS red blood cell count trait[138] that were also found significantly associated with whole blood expression factors identified in a separate study (unpublished).

We applied GSFA to this dataset with the hope of identifying genes or gene modules affected by risk variants for the aforementioned complex immune trait. Unfortunately, no significant DEGs (LFSR < 0.1) were detected under any of the SNPs. Despite specifying 50 factors in total, we only found significant effects (PIP > 0.95) of 3 SNPs on factor 1; 8 SNPs are associated with 4 factors with PIP > 0.5 (Fig. 3.12a). We also tested the association between each pair of GSFA inferred factor and SNP condition using simple linear regression, but no significant association was detected at FDR < 0.1 (Fig. 3.12b). Because of the two-component normal-mixture prior imposed on gene weights, the model did not identify any "effective" foreground genes apart from the background for factors with inherently dense and symmetric gene weights (factors 1-3, Fig. 3.12e); for other factors with less symmetrically distributed gene weights, genes in the longer tail tend to be identified as the foreground (Fig. 3.12f). At a cutoff of gene PIP > 0.9, the density of loaded genes for each GSFA factor ranges from 0 to 13.4%. GO enrichment of loaded genes in factors found significant gene sets in 34 of the 50 factors. For example, factor 26, the only other factor somewhat associated with a SNP condition (PIP = 0.56 and effect size = -0.37 with the chr16 SNP), has the most number (128) of enriched gene sets among all factors, and these enriched gene sets belong to a variety of categories, from immune response to cell development (Fig. 3.12d).

Alternatively, we also applied a two-step approach, where we first inferred expression factors using another Bayesian sparse factor analysis model, FLASH[118], and then associated them with

each SNP condition via simple linear regression. The result is similar to that of GSFA, with no significant associations detected at FDR < 0.1 (Fig. 3.12c). While FLASH uses a different shrinkage prior (Equation 9), the first 3 factors inferred by FLASH and by GSFA are highly correlated (Pearson R = 0.86-0.99), and have similar density of gene loadings on them, consistent with the common observation that it is hard to induce sparsity in the top factors (ranked by percentage variance explained) from bulk gene expression data, which are inherently dense[118]. An alternative way to interpret these dense factors, which often contain rich information, would be to conduct gene set enrichment analysis using the top genes ranked by gene weights in both positive and negative directions, which is commonly used for interpretation of principal components from PCA. However, even if significant associations were found, this would be an indirect two-step approach to link perturbations with specific downstream genes. Overall, this case study highlighted the difficulty when applying Bayesian sparse factor models to bulk RNA-seq data.

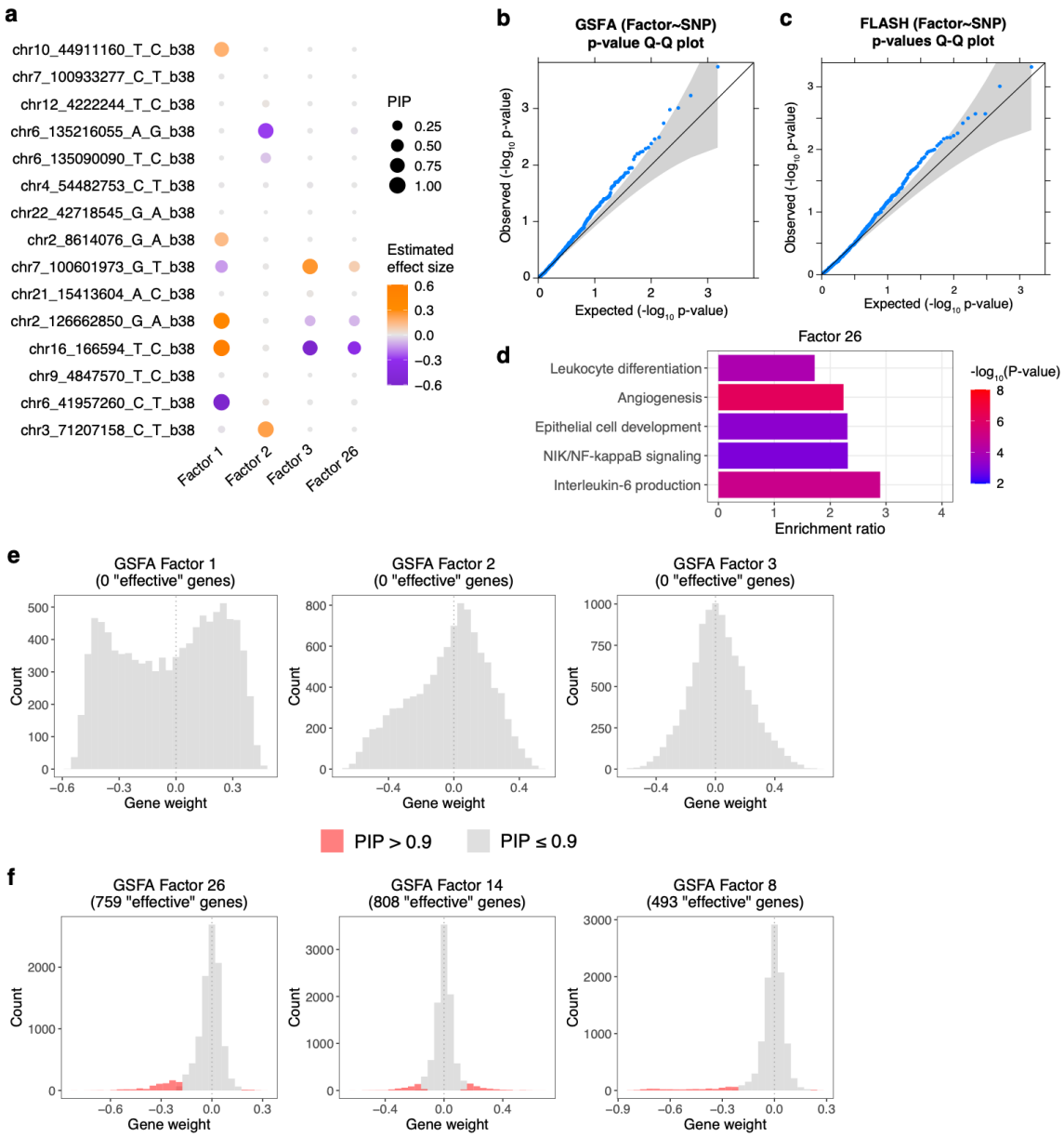

Figure 3.12: GSFA and FLASH results on GTEx whole blood bulk RNA-seq dataset.
a) Estimated effects of gene perturbations on factors inferred by GSFA with at least one association PIP > 0.5. The size of a dot represents the PIP of association; the color represents the effect size. **b**), **c**) Q-Q plot of p-values from simple linear regression between each pair of SNP genotype and factor inferred by GSFA (b) or FLASH (c). **d**) The fold of enrichment for selected GO "biological process" terms enriched in GSFA factor 26 (q-value < 0.05). Hypergeometric test was used, where genes with PIP > 0.9 in the factor were compared against a background of all 10k genes used in GSFA. **e**), **f**) Distribution of GSFA inferred gene weights in factors 1-3 (e), and in factors 26, 14, 8 (f). Genes that have a gene PIP > 0.9 were colored in red; factors 1-3 do not have any significant genes at this cutoff.

3.4.6 Additional methods

Deviance residual transformation of gene count data for GSFA application

The GSFA model assumes that the expression matrix **Y** is normally distributed. To accommodate the application of GSFA on count data, we follow the transformation proposed in Townes *et al*.[26], where the count data are transformed into continuous quantities in the form of deviance residuals. In a standard data normalization pipeline, raw counts are normalized by sample-specific size factors, and then log-transformed. However, due to the large number of zeros in scRNA-seq UMI counts, normalization schemes commonly used for bulk RNA-seq data may result in unstable normalization[139], and the arbitrary pseudocount added during the log transformation of exact zeros may distort the data, introducing systematic errors and causing spurious differences in expression[140]. The deviance residual transformation circumvents these difficulties by directly modeling the raw count data under a multinomial null model of constant gene expression across all cells, and quantifying the fit of data in the form of deviance residuals, a quantity analogous to z-scores and approximately follow a normal distribution. Specifically, the deviance residual for gene *j* in cell *i* is

$$r_{ij} = \text{sign}\left(c_{ij} - \hat{\mu}_{ij}\right)\sqrt{2c_{ij}\log\frac{c_{ij}}{\hat{\mu}_{ij}} + 2(n_i - c_{ij})\log\frac{n_i - c_{ij}}{n_i - \hat{\mu}_{ij}}}.$$

Here, c_{ij} is the raw gene count, n_i is the library size of cell *i*, and $\hat{\mu}_{ij} = n_i\frac{\sum_i c_{ij}}{\sum_i n_i}$ is the expression of gene *j* under the null model of constant expression.

Following Townes *et al*.[26], we use an approximate multinomial deviance statistic to evaluate the deviance of each gene from the null model:

$$D_j = \sum_{i=1}^{N} r_{ij}^2$$

Genes with constant expression across cells are not informative and will have a deviance of 0, while genes that vary across cells in expression will have a larger deviance. Therefore, when selecting informative features for downstream analysis, one can pick the genes with high deviance as an alternative to selecting highly variable genes, with the advantage that the selection is not sensitive to normalization.

Alternative differential gene expression (DGE) methods

For comparison, we applied the following DGE methods to simulated or real data:

(1) Welch's t-test[126] (two-sided) using the t.test() function in R;

(2) edgeR quasi-likelihood F-test (edgeR-QLF)[73] using glmQLFit() and glmQLFTest() functions in the R package edgeR;

(3) DESeq2[90] using the DESeq() function in the R package DESeq2;

(4) MAST[91], a statistical method tailored for scRNA-seq data, using zlm() and lrTest() functions in the R package MAST;

(5) scMAGeCK-LR[27], a linear-regression-based approach tailored for single-cell CRISPR screening data, using the scmageck_lr() function in the R package scMAGeCK. We did not include scMAGeCK-RRA as it is not designed to test all genes[27].

GSFA analysis of $CD8^+$ T cell CROP-seq dataset

Raw cellranger outputs of the $CD8^+$ T cell CROP-seq study[95] were downloaded from Gene Expression Omnibus (GEO: GSE119450). We merged resting and stimulated T cells from two donors using the R package Seurat_4.0.1[141]. We first filtered cells that contain fewer than 500 expressed genes or more than 10% of total read counts from mitochondria genes, keeping 14278 stimulate T cells and 10677 unstimulated T cells. Next, we transformed the raw counts into deviance residuals for all genes in all cells, kept the top 6000 genes ranked by deviance statistics,

then regressed out unique UMI count, library size and percentage of mitochondrial gene expression from the reduced deviance residual matrix. The resulting matrix was then scaled so that each gene has variance 1.

The gRNA perturbation data are binarized, with gRNAs targeting the same gene deemed as the same type of perturbation. The scaled gene expression matrix and the perturbation matrix were used as input for GSFA. To capture potentially different effects of CRISPR perturbation under resting and stimulated conditions, we used the modified GSFA model with two cell groups (Section 3.3.4), stratifying all cells by their stimulation states (unstimulated:0, stimulated:1). We specified 20 factors in the model. Gibbs sampling was performed for 4000 iterations, and posterior means of parameters were computed from the last 1000 iterations. Computation took ~11 hours on a modern Linux workstation with Intel Xeon E5-2680 v4 (2.40 GHz) processors.

We assessed the calibration of the GSFA results using permutation. We created 10 permutation sets on the stimulated and the unstimulated cells, separately. In each permutation set, the cell labels were permuted independently of the perturbation conditions, and GSFA was run on each of these datasets. The calibration was assessed in a few ways. We checked the distribution of PIPs of the perturbation effects on factors ($\boldsymbol{\beta}$), and the distribution of LSFRs from the inferred perturbation to gene effects. We expect PIPs to be close to 0 and LSFRs close to 1 in the permutation results. We also assessed the empirical p-values of correlations between perturbations and inferred factors. Since we do not expect any correlation between the two under permutation, any deviation of p-values from the null distribution would indicate that GSFA incorrectly borrows information from perturbations to infer factors, a potential problem that would inflate the results. All permutation results are reported in Fig. 3.8.

GSFA analysis of LUHMES CROP-seq dataset

Raw cellranger outputs of the LUHMES neural progenitor cell CROP-seq study[135] were downloaded from Gene Expression Omnibus (GEO: GSE142078). We merged all 3 batches of LUHMES CROP-seq raw data together using the R package Seurat_4.0.1[141], and filtered cells with a library size over 20000 or more than 10% of total read counts from mitochondria genes, keeping 8708 cells. Similarly, we transformed the raw count matrix into a reduced deviance residual matrix with top 6000 genes ranked by deviance residual. Differences in experimental batch, unique UMI count, library size, and percentage of mitochondrial gene expression were all regressed out. Running of GSFA is similar to before, except that there is only one cell group, and that Gibbs sampling was run for 3000 iterations, which took ~2.5 hours on a modern Linux workstation with Intel Xeon E5-2680 v4 (2.40 GHz) processors. We also assessed calibration of GSFA results, in the same way as we did in the T cell analysis. The results are reported in Fig. 3.11.

Running alternative DGE methods on CD8$^+$ T cell and LUHMES CROP-seq data

For both stimulated T cells and LUHMES CROP-seq data, we performed alternative DGE analyses for comparison. We applied edgeR-QLF[73], DESeq2[90], and MAST[91] directly to the scRNA-seq raw count data, contrasting cells with each perturbation from those without, for all the genes selected for GSFA. For the LUHMES dataset, experimental batch was included as one of the covariates in these 3 tests. We also applied scMAGeCK-LR[27] to the transformed and corrected CROP-seq data (described above). For all these methods, FDR was computed following the Benjamini-Hochberg procedure for genes under each perturbation, and significant DEGs were obtained under an FDR cutoff of 0.05.

To assess the calibration of the differential expression test p-values from these methods, we carried out permutation tests for each DGE method by randomly shuffling the cell labels

independent of the perturbation conditions. For the T cell dataset, shuffling occurred within the stimulated cells. We generated 10 permuted datasets, and performed the DGE methods in the same way as before.

Gene ontology enrichment analysis

Gene ontology (GO) over-representation analyses were performed using the WebGestaltR() function in the R package WebGestaltR_0.4.4[142] with default parameters and the functional category for enrichment analysis set to the GO Slim "Biological Process" category (geneontology_Biological_Process_noRedundant). To interpret GSFA inferred factors (gene modules), genes with weight PIP > 0.95 were treated as the foreground, while all genes used in GSFA were treated as the background in the over-representation analysis. To interpret DEGs discovered under each perturbation by GSFA or other DGE methods, genes with LSFR < 0.05 (or FDR < 0.05) were treated as the foreground, while all genes evaluated were treated as the background in the over-representation analysis.

GSFA analysis of GTEx bulk RNA-seq data

GTEx whole blood RNA-seq data from 670 individuals and paired genotype data were obtained from the GTEx Portal and dbGaP v8 on 2017-06-05. The 15 GWAS red blood cell count risk SNPs used as genetic perturbations were selected based on a separate unpublished study. Briefly, these 15 SNPs are GWAS risk SNPs (after p-value thresholding and LD pruning) that were significantly associated with expression factors from the whole blood data. Expression factors was obtained using PLIER[107] constrained by a prior matrix comprised of immune and chemgen pathways from MSigDB[143].

Genotypes were converted to numbers according to the additive model "0/0": 0, "0/1": 1, "1/1": 2. Genotypes at these 15 SNPs were matched with their corresponding individual expression

profiles in whole blood. 647 individuals were matched with valid genotypes at all these SNPs. TPM-normalized gene expression data was corrected for the first 5 genotype PCs, PCR, platform, and sex covariates. Top 10k most variable genes were used in downstream analyses. GSFA was performed with the reduced normalized gene expression matrix and the matched SNP genotype matrix as inputs. Specifying 50 factors in the model, Gibbs sampling was run for 3000 iteration. We also performed FLASH[118] on the reduced normalized gene expression matrix using R package flashier_0.2.7 with default parameters and 50 specified factors.

3.5 Discussion

Single-cell CRISPR screening technologies have enabled efficient readouts of transcriptome-level effects of multiple genetic perturbations in tens of thousands of individual cells in a single experiment. These technologies offer great opportunities, but also challenges for effective data analysis. We presented GSFA to address these challenges. GSFA can identify biologically interpretable gene modules that respond to genetic perturbations, and by summarizing the information from all factors, GSFA can infer the effects of perturbations on specific downstream genes. When applied to two CROP-seq datasets, GSFA detected transcriptomic effects and shed light on the molecular mechanisms of regulators of T cell activation and neuronal differentiation, respectively.

GSFA is built on factor analysis, which we believe offers key benefits over clustering-based analysis of single-cell screening data[144], [145]. Clustering-based analysis requires the effects of perturbations to be large enough to alter cluster compositions; thus, it may miss perturbations with moderate effects. In contrast, factor analysis does not rely on disjoint cell clusters and can potentially detect subtler effects, *e.g.* those of non-coding regulatory elements. In addition, as we have demonstrated, the inferred factors lead to better biological interpretability than cell clusters.

Conceptually, a researcher can perform factor analysis or related methods such as topic modeling on the expression data, and then correlate the inferred factors with genetic perturbations across cells. This approach was used in the MUSIC method[146]. Compared with this two-step approach, GSFA has several advantages. When inferring expression factors, GSFA uses the genetic perturbation as a prior to improve the estimation of the factors (hence "guided" in the method name). In practice, GSFA offers an important advantage when a perturbation affects multiple factors. For example, in the LUHMES data, target genes often show associations with two or more factors (Fig. 3.9a), which may have overlapping functions (Fig. 3.9d). Thus, a perturbation may have positive effects on some factors, and negative effects on others. Similarly, some factors may correlate positively with the biological process they represent, *e.g.* cell maturation, while others correlate negatively. In such cases, it would be extremely difficult to learn the total effects of the perturbation. GSFA solves this problem by synthesizing the effects of a perturbation over all factors.

In GSFA, factors are dependent on the genetic perturbations via a linear model. In this sense, GSFA is related to a class of factor models in the statistics literature, sometimes called supervised factor analysis, where the factors depend on covariates of samples[147]–[149]. These models can help improve the estimation of latent factors, and have been proposed in bulk gene expression data analysis[150] where samples have different characteristics or experimental conditions. Nevertheless, existing covariate-dependent factor models were designed only for factor inference, and do not provide estimates of the effects of covariates (perturbations in our case) for specific genes, *i.e.*, they cannot estimate perturbation effects for single-cell CRISPR screening data.

GSFA is a versatile tool for single-cell screening data analysis and can be potentially used in other settings. As demonstrated in our study of $CD8^+$ T cell data, GSFA can be used in datasets

with multiple cell types or treatment conditions, and estimate the perturbation effects for each cell group separately. Secondly, while both applications used the low multiplicity of infection (MOI) setting, with each cell typically harboring at most one gRNA perturbation, GSFA's linear model of how factors depend on genetic perturbations easily accommodates multiple perturbations per cell, and thus can be readily applied to high MOI settings. Thirdly, the implementation of GSFA in Rcpp makes it relatively efficient, and its distribution in the form of an R package makes it accessible and easily incorporated into common data analysis of single-cell RNA-seq in R.

In principle, the generality of the statistical model of GSFA makes it applicable to bulk RNA-seq studies with paired genetic perturbations. This has been demonstrated in its application on GTEx expression data with paired genotype measurements, but with insignificant results. This could be due to several reasons. Because GTEx bulk tissue RNA-seq data is a collective measurement over multiple cell types, this may have complicated the original modular patterns in the expression data, thus limiting the detection of sparse gene modules. In addition, since genetic variants tend to have cell-type-specific effects, any association of them with inferred gene modules would be even harder to detect. All of these highlight the advantages of single-cell RNA-seq in studying genetic effects.

GSFA can be further improved along several directions. GSFA does not directly model read counts and instead uses deviance residuals converted from count data. While this transformation generally works well in our analysis, we noticed that LFSRs from differential expression analysis can be modestly inflated at high factor density (Fig. 3.3b, under $\pi = 0.2$). Hence, directly modeling the read count data may improve the calibration and power of GSFA. Another limitation of GSFA is that we assume genetic perturbations affect downstream genes through factors. It is possible that the factors may not fully capture the transcriptional effects; thus, it may be desirable to add to the

model “direct effect" terms, where perturbations directly affect the expression of a gene without acting on any factors. Finally, GSFA uses Gibbs sampling for inference; replacing this with a more efficient algorithm, such as variational approximation, may reduce the computational time.

In conclusion, single cell CRISPR-screening is a promising technology, yet the difficulty of data analysis has prevented us from realizing its full potential. GSFA complements the strength of this technology by offering a powerful new analysis framework, representing a substantial advance of the field.

CHAPTER 4: CONCLUSION

The work in this dissertation evolves around an overarching goal: the development and integration of computational frameworks to understand the regulatory effects of genetic variants, thereby aiding the genetic discovery of diseases.

In Chapter 1, I developed a preliminary computational framework to infer gene regulatory networks from matching chromatin accessibility and gene expression data. Using ATAC-seq data of iPSC and derived neuronal cell types, I mapped open chromatin regions that indicate *cis*-regulatory elements (CREs), and inferred transcription factor (TF) binding in CREs through TF footprint calling. Using matched RNA-seq data, I further identified CREs with accessibility levels associated with putative target gene expression, forming a connection from TFs and CREs to their regulatory gene target. I inferred such regulatory networks for 85 high-confidence autism risk genes within our neurodevelopmental model, identifying 849 enhancer regions with regulatory potential on their expression levels. These enhancer regions, at the meantime, are enriched for predicted binding sites of important TFs involved in neural development or differentiation. Hence, we have prioritized non-coding regions that likely regulate ASD risk genes during brain development, providing mechanistic insights that are potentially useful for the interpretation of genetic variants associated with ASD susceptibility.

The mapping of enhancers and their target genes has always been a challenge, particularly since an enhancer can act independently of its orientation and distance from the transcription starting site[30]. Our approach provides an efficient way to link genes to their putative distal enhancers genome wide in any system given matched gene expression and accessibility data. However, the inferred connections should be treated as preliminary results since our prediction is based on simplified assumptions: we approximated enhancer activity using chromatin accessibility,

and assumed that functional enhancers have their activity levels correlated with the target gene expression in a linear fashion. While enhancer accessibility is somewhat indicative of its activity, it can be influenced by other factors such as their TF binding states and TF concentrations. Although it is possible to infer the general TF binding states of these CREs, how the composition of TFBSs affect gene expression outcome is not well-explored; in addition, a collective of enhancers can affect gene expression in non-linear ways such as synergistically or binarily[151]. Therefore, experimental follow-ups such as luciferase reporter assay and CRISPR editing are warranted to validate the regulatory functions of these identified CREs and their inferred target genes.

With regulatory sequences, or more generally, the open chromatin regions (OCRs) of the genome identified, we ultimately hope that they can help us interpret causal disease variants. However, since not all variants in OCRs are functional, it remains a challenge to pinpoint the causal functional variants for complex traits/diseases such as schizophrenia (SZ). In Chapter 2, we provided a novel framework to address this challenge, leveraging allelic-specific open chromatin (ASoC) information. We created the first snapshot of ASoC landscape in iPSC and derived neurons. We found that a large fraction of ASoC variants are cell-type-specific, with neuronal ASoC variants enriched in expression and methylation QTLs in brain and partially driven by alternation of TF binding. We also found strong enrichment of SZ GWAS risk variants in these neuronal ASoC variants, leveraging which we were able to identify putative causal variants of SZ. Using multiplexed single-cell CRISPRi screening followed by independent CRISPR editing, we confirmed the functional effects at 6 SZ risk variant loci and identified 9 corresponding *cis*-target genes. Together, this study demonstrated that ASoC data in iPSC-derived neurons provide an

effective means to interpret functional effects of genetic variants in the neurodevelopment context, and to aid genetic discovery of brain-related traits.

Although genetic variants associated with neuropsychiatric disorders have been found enriched in OCRs identified from postmortem brains, whether these genetic risk variants affect chromatin accessibility during neurodevelopment had not been demonstrated. Our study not only provided a useful resource to chase down causal variants, but also generated direct evidence that neuropsychiatric risk variants alter chromatin accessibility during neurodevelopment, which is likely one of the main mechanisms of these disorders. We envision similar approaches can be applied to other biological systems to aid the discovery of causal regulatory variants for other complex diseases/traits.

Our ASoC approach for variant discovery is in the same spirit as chromatin accessibility QTL (caQTL) studies. ASoC contrasts the accessibility between two alleles at a heterozygous site within the same individual (or in our cases, allelic read counts pooled over heterozygous individuals), so the sample size might be limited. caQTL studies, on the other hand, compare the accessibility data across individuals, and thus, tend to have higher discovery power. In both cases, careful measures are needed when processing and analyzing allelic-specific signals, as they are subject to biases such as genotyping errors, imprinting, reference mapping bias, and PCR amplification bias[152]. In addition, when data from multiple individuals are utilized, the binomial assumption on allele-specific read counts may be over-simplified, and more sophisticated statistical methods such as RASQUAL[152] are needed to capture the over-dispersion in read counts due to between-individual variations.

While alteration of chromatin accessibility is one major mechanism causal variants act through, we do realize that the focus on ASoC limits the scope of discovery, as functional variants

can also affect downstream phenotypes through post-transcriptional mechanisms such as alternative splicing[92] and mRNA modification[153]. These additional functional annotations, combined with statistical fine-mapping[60], would further improve the discovery of causal disease variants.

We evaluated the functions of prioritized SZ risk SNPs and their *cis*-regulating genes using high-throughput single-cell CRISPRi screens. With routine differential expression (DE) analysis methods, however, we were only able to detect a limited number of *cis*-genes for the targeted loci, and no *trans*-effects. One way to boost the detection power is to increase the sample size, *i.e.*, the number of cells subject to each genetic perturbation. This can be achieved through either including more cells in the experiment, or switching to a high multiplicity of infection (MOI)[23] CRISPR setting. Computationally, instead of traditional DE methods where genes are tested one at a time, alternative statistical methods that leverage the modular structure in gene expression may be able to increase the discovery power of the subtle but often coordinated effects on genes imparted by these genetic perturbations.

In Chapter 3, I developed such a novel statistical framework, GSFA, to effectively analyze the transcriptomic effects of genetic perturbation from high-throughput single-cell CRISPR screening data. Built upon a factor analysis model, GSFA can extract from gene expression data coordinated gene modules (factors) that affected the perturbations. And by synthesizing the effects of a perturbation over all factors, GSFA can infer the DE effects on specific downstream genes. When applied to two CROP-seq datasets, GSFA identified biologically relevant gene modules, and had better power to detect differentially expressed genes than alternative methods, shedding light on the molecular mechanisms regulating T cell activation and neuronal differentiation.

The idea of leveraging co-regulated gene models has been applied to single-cell CRISPR screening data in the form of clustering or dimensional reduction methods followed by association of clusters[95], [96] or lower dimensional structures[22], [135], [146] with the perturbation condition. In contrast to the cell clusters or dense lower-dimensional components recovered by these methods, factors inferred by GSFA are more interpretable thanks to the sparse priors imposed on the gene loadings. During the inference, GSFA incorporated the genetic perturbation(s) as a prior, which could improve the estimation of the factors when the perturbations have large effects. More importantly, GSFA offers an important advantage over these two-step approaches in its ability to summarize a total effect of given genetic perturbation on each specific downstream gene with ease, especially when the perturbation affects multiple factors with overlapping functions.

One major assumption of GSFA is that genetic perturbations affect a broad set of downstream genes through factors. Accordingly, GSFA was primarily designed for experimental settings that target important regulatory genes, such as transcription factors and signaling proteins, which are well-known for having broad effects once perturbed. However, in cases where enhancers or genetic variants are targeted and the perturbation effects are relatively small, especially at the transcriptome level, we do not expect GSFA to be more advantageous than routine DE methods, despite that the problem which originally inspired GSFA involves the targeting of putative enhancers. Since GSFA may not capture transcriptional effects that are not mediated through factors, it may be desirable to add to the model a “direct effect" term, where perturbations directly affect the expression of a gene without acting on any factors. Another limitation of GSFA is, it operates on normalized expression data rather than directly modeling the count data. In the study, we used deviance residual transformation as a preprocessing step to remove technical factors.

While this transformation has proven to work well for single-cell UMI count data[26], directly modeling the count data may improve the calibration and power of GSFA.

High-throughput single-cell CRISPR screens such as CROP-seq[22], Perturb-seq[20], [21], MOSAIC-seq[154] etc., are among the most promising technologies to identify functional CREs and their gene targets across the genome. However, careful measures are needed to calibrate DE tests and validate the results, especially since single-cell CRISPR screen readouts are subject to many potential confounding factors. For example, in high MOI experiments, the presence of gRNA perturbation in a cell is confounded by technical factors such as sequencing depth and batch effect. This can be addressed by modeling gRNA presence in cells based on technical factors followed by conditional resampling, as proposed by SCEPTRE[28]. The prevalence of CRISPR off-target effects[155] is another challenge that limits the specificity of DE analyses. Therefore, gRNA sequences need to be carefully designed to lower the likelihood of off-target cutting[156]; it is also beneficial to target the same locus with multiple gRNAs, and use the concordance among DE genes of these gRNAs as an estimate of on/off-target effects[157]. For non-coding CRISPR screens, it is less well-explored what causes the disruption of enhancer function, and therefore, it may be necessary to either introduce larger deletions[158] or more complete tiling[159] for the region of interest to increase the effect size of perturbation.

Similar to what has been demonstrated in Chapter 2, pooled single-cell CRISPR screens can aid the discovery of GWAS functional variants, with all loci of putative causal variants obtained from statistical fine-mapping being perturbed at scale in their native genomic contexts and their effects being directly measured via single-cell transcriptome readout. Different CRISPR perturbation methods from CRISPR-Cas9 nuclease cut, CRISPR-Cas9 interference (CRISPRi), to CRISPR-Cas9 activation (CRISPRa)[16] offer a variety of venues to derive novel regulatory

insights. While powerful, single-cell CRISPR screens, especially those targeting non-coding sequences, still suffer from experimental and computational challenges as discussed above. Therefore, the discoveries of enhancers and their target genes need to be evaluated together with orthogonal evidence such as expression QTL, ChIP-seq, and promoter capture Hi-C data.

Besides transcriptome readout, single-cell CRISPR genetic screens can also be used to detect perturbation effects on chromatin state. Spear-ATAC[160] and CRISPR-sciATAC[161], recent protocols that combine multiplexed CRISPR with ATAC-seq in single cells in a high-throughput manner, can directly link genetic perturbations at key TFs or chromatin remodelers to genome-wide chromatin accessibility changes. It is also possible to characterize the effect of genetic perturbation on phenotypes downstream of gene expression with CRISPR screens. For example, ECCITE-seq[162] can interrogate the changes in expression of surface protein markers in addition to the transcriptome of each single cell. Since the measurement of proteins have low dropout, the additional information can help confirm or even improve the discovery of differentially expression phenotypes responding to the perturbation, especially in immune cells[162].

Taken together, advances in both experimental technologies and computational methods have greatly improved our ability to understand the effects of genetic variants and their link to disease phenotypes. From mapping of genome-wide regulatory features to prioritization of GWAS disease variants using statistical fine-mapping, and from functional validation of these prioritized risk loci using high-throughput CRIPSR screens to statistical analyses for dimensionality reduction and differential expression tests, key genetic variants for complex diseases and their effects on downstream molecular phenotypes are gradually unraveled. I look forward to a future where more and more causal genetic variants and their disease-causing mechanisms come into light.

REFERENCES

[1] E. S. Lander *et al.*, "Initial sequencing and analysis of the human genome," *Nat. 2001 4096822*, vol. 409, no. 6822, pp. 860–921, Feb. 2001, doi: 10.1038/35057062.

[2] I. Dunham *et al.*, "An integrated encyclopedia of DNA elements in the human genome," *Nature*, vol. 489, no. 7414, pp. 57–74, 2012, doi: 10.1038/nature11247.

[3] S. L. Klemm, Z. Shipony, and W. J. Greenleaf, "Chromatin accessibility and the regulatory epigenome," *Nat. Rev. Genet. 2018 204*, vol. 20, no. 4, pp. 207–220, Jan. 2019, doi: 10.1038/s41576-018-0089-8.

[4] L. Song and G. E. Crawford, "DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells," *Cold Spring Harb. Protoc.*, vol. 2010, no. 2, p. pdb.prot5384, Feb. 2010, doi: 10.1101/PDB.PROT5384.

[5] J. D. Buenrostro, B. Wu, H. Y. Chang, and W. J. Greenleaf, "ATAC-seq: A method for assaying chromatin accessibility genome-wide," *Curr. Protoc. Mol. Biol.*, vol. 2015, pp. 21.29.1-21.29.9, 2015, doi: 10.1002/0471142727.mb2129s109.

[6] D. Shlyueva, G. Stampfel, and A. Stark, "Transcriptional enhancers: From properties to genome-wide predictions," *Nat. Rev. Genet.*, vol. 15, no. 4, pp. 272–286, 2014, doi: 10.1038/nrg3682.

[7] A. Rada-Iglesias, R. Bajpai, T. Swigut, S. A. Brugmann, R. A. Flynn, and J. Wysocka, "A unique chromatin signature uncovers early developmental enhancers in humans," *Nat. 2010 4707333*, vol. 470, no. 7333, pp. 279–283, Dec. 2010, doi: 10.1038/nature09692.

[8] D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold, "Genome-wide mapping of in vivo protein-DNA interactions," *Science (80-.).*, vol. 316, no. 5830, pp. 1497–1502, Jun. 2007, doi: 10.1126/SCIENCE.1141319/SUPPL_FILE/JOHNSON.SOM-5-30.PDF.

[9] S. Neph *et al.*, "An expansive human regulatory lexicon encoded in transcription factor footprints," *Nat. 2012 4897414*, vol. 489, no. 7414, pp. 83–90, Sep. 2012, doi: 10.1038/nature11212.

[10] S. S. P. Rao *et al.*, "A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping," *Cell*, vol. 159, no. 7, pp. 1665–1680, 2014, doi: 10.1016/j.cell.2014.11.021.

[11] J. Dekker, K. Rippe, M. Dekker, and N. Kleckner, "Capturing Chromosome Conformation," *Science (80-.).*, vol. 295, no. 5558, pp. 1306–1311, Feb. 2002, doi: 10.1126/SCIENCE.1067799.

[12] B. Mifsud *et al.*, "Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C," *Nat. Genet. 2015 476*, vol. 47, no. 6, pp. 598–606, May 2015, doi: 10.1038/ng.3286.

[13] A. Buniello *et al.*, "The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D1005–D1012, Jan. 2019, doi: 10.1093/NAR/GKY1120.

[14] M. T. Maurano *et al.*, "Systematic Localization of Common Disease-Associate Variation in Regulatorty DNA," *Science (80-.).*, vol. 337, no. September, pp. 1190–1195, 2012, doi: 10.1126/science.1222794.

[15] M. Jinek, K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna, and E. Charpentier, "A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity," *Science (80-.).*, vol. 337, no. 6096, pp. 816–821, Aug. 2012, doi: 10.1126/science.1225829.

[16] M. Gasperini, J. M. Tome, and J. Shendure, "Towards a comprehensive catalogue of validated and target-linked human enhancers," *Nat. Rev. Genet. 2020 215*, vol. 21, no. 5, pp. 292–310, Jan. 2020, doi: 10.1038/s41576-019-0209-0.

[17] R. Lopes, G. Korkmaz, and R. Agami, "Applying CRISPR–Cas9 tools to identify and characterize transcriptional enhancers," *Nat. Rev. Mol. Cell Biol. 2016 179*, vol. 17, no. 9, pp. 597–604, Jul. 2016, doi: 10.1038/nrm.2016.79.

[18] X. Gao, J. C. H. Tsang, F. Gaba, D. Wu, L. Lu, and P. Liu, "Comparison of TALE designer transcription factors and the CRISPR/dCas9 in regulation of gene expression by targeting enhancers," *Nucleic Acids Res.*, vol. 42, no. 20, p. e155, Nov. 2014, doi: 10.1093/NAR/GKU836.

[19] G. Korkmaz *et al.*, "Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9," *Nat. Biotechnol. 2015 342*, vol. 34, no. 2, pp. 192–198, Jan. 2016, doi: 10.1038/nbt.3450.

[20] A. Dixit *et al.*, "Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens," *Cell*, vol. 167, no. 7, pp. 1853-1866.e17, Dec. 2016, doi: 10.1016/j.cell.2016.11.038.

[21] B. Adamson *et al.*, "A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response," *Cell*, vol. 167, no. 7, pp. 1867-1882.e21, Dec. 2016, doi: 10.1016/J.CELL.2016.11.048.

[22] P. Datlinger *et al.*, "Pooled CRISPR screening with single-cell transcriptome readout," *Nat. Methods*, vol. 14, no. 3, pp. 297–301, Jan. 2017, doi: 10.1038/nmeth.4177.

[23] M. Gasperini *et al.*, "A Genome-wide framework for mapping gene regulation via cellular genetic screens," *Cell*, vol. 176, no. 1–2, pp. 377–390, Jan. 2019, doi: 10.1016/J.CELL.2018.11.029.

[24] J. L. McFaline-Figueroa, A. J. Hill, X. Qiu, D. Jackson, J. Shendure, and C. Trapnell, "A pooled single-cell genetic screen identifies regulatory checkpoints in the continuum of the epithelial-to-mesenchymal transition," *Nat. Genet.*, vol. 51, no. 9, pp. 1389–1398, Sep.

2019, doi: 10.1038/s41588-019-0489-5.

[25] X. Jin *et al.*, "In vivo Perturb-Seq reveals neuronal and glial abnormalities associated with autism risk genes," *Science*, vol. 370, no. 6520, p. eaaz6063, Nov. 2020, doi: 10.1126/SCIENCE.AAZ6063.

[26] F. W. Townes, S. C. Hicks, M. J. Aryee, and R. A. Irizarry, "Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model," *Genome Biol.*, vol. 20, no. 1, pp. 1–16, Dec. 2019, doi: 10.1186/S13059-019-1861-6/FIGURES/5.

[27] L. Yang *et al.*, "ScMAGeCK links genotypes with multiple phenotypes in single-cell CRISPR screens," *Genome Biol.*, vol. 21, no. 1, pp. 1–14, Jan. 2020, doi: 10.1186/S13059-020-1928-4/FIGURES/5.

[28] T. Barry, X. Wang, J. A. Morris, K. Roeder, and E. Katsevich, "SCEPTRE improves calibration and sensitivity in single-cell CRISPR screen analysis," *Genome Biol.*, vol. 22, no. 1, pp. 1–19, Dec. 2021, doi: 10.1186/S13059-021-02545-2/FIGURES/5.

[29] T. Wang, B. Li, C. E. Nelson, and S. Nabavi, "Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data," *BMC Bioinformatics*, vol. 20, no. 1, pp. 1–16, Jan. 2019, doi: 10.1186/S12859-019-2599-6/TABLES/7.

[30] J. Banerji, S. Rusconi, and W. Schaffner, "Expression of a β-globin gene is enhanced by remote SV40 DNA sequences," *Cell*, vol. 27, no. 2, pp. 299–308, Dec. 1981, doi: 10.1016/0092-8674(81)90413-X.

[31] R. Elkon and R. Agami, "Characterization of noncoding regulatory DNA in the human genome," *Nat. Biotechnol.*, vol. 35, no. 8, pp. 732–746, 2017, doi: 10.1038/nbt.3863.

[32] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome Biol.*, vol. 10, no. 3, pp. 1–10, Mar. 2009, doi: 10.1186/GB-2009-10-3-R25/TABLES/5.

[33] H. Li *et al.*, "The Sequence Alignment/Map format and SAMtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, Aug. 2009, doi: 10.1093/BIOINFORMATICS/BTP352.

[34] K. D. Hansen, R. A. Irizarry, and Z. Wu, "Removing technical variability in RNA-seq data using conditional quantile normalization," *Biostatistics*, vol. 13, no. 2, p. 204, Apr. 2012, doi: 10.1093/BIOSTATISTICS/KXR054.

[35] A. Dobin *et al.*, "STAR: ultrafast universal RNA-seq aligner," *Bioinformatics*, vol. 29, no. 1, pp. 15–21, Jan. 2013, doi: 10.1093/BIOINFORMATICS/BTS635.

[36] M. H. Sung, M. J. Guertin, S. Baek, and G. L. Hager, "DNase Footprint Signatures are Dictated by Factor Dynamics and DNA Sequence," *Mol. Cell*, vol. 56, no. 2, p. 275, Oct. 2014, doi: 10.1016/J.MOLCEL.2014.08.016.

[37] R. Pique-Regi, J. F. Degner, A. A. Pai, D. J. Gaffney, Y. Gilad, and J. K. Pritchard,

"Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data," *Genome Res.*, vol. 21, no. 3, pp. 447–455, Mar. 2011, doi: 10.1101/GR.112623.110.

[38] Z. Li, M. H. Schulz, T. Look, M. Begemann, M. Zenke, and I. G. Costa, "Identification of transcription factor binding sites using ATAC-seq," *Genome Biol.*, vol. 20, no. 1, Feb. 2019, doi: 10.1186/s13059-019-1642-2.

[39] E. G. Gusmao, C. Dieterich, M. Zenke, and I. G. Costa, "Detection of active transcription factor binding sites with the combination of DNase hypersensitivity and histone modifications," *Bioinformatics*, vol. 30, no. 22, pp. 3143–3151, Nov. 2014, doi: 10.1093/BIOINFORMATICS/BTU519.

[40] M. P. Forrest *et al.*, "Open Chromatin Profiling in hiPSC-Derived Neurons Prioritizes Functional Noncoding Psychiatric Risk Variants and Highlights Neurodevelopmental Loci," *Cell Stem Cell*, vol. 21, no. 3, p. 305, Sep. 2017, doi: 10.1016/J.STEM.2017.07.008.

[41] K. S. O'Shea and M. G. McInnis, "Neurodevelopmental origins of bipolar disorder: iPSC models," *Mol. Cell. Neurosci.*, vol. 73, pp. 63–83, Jun. 2016, doi: 10.1016/J.MCN.2015.11.006.

[42] B. S. Abrahams *et al.*, "SFARI Gene 2.0: A community-driven knowledgebase for the autism spectrum disorders (ASDs)," *Mol. Autism*, vol. 4, no. 1, pp. 1–3, Oct. 2013, doi: 10.1186/2040-2392-4-36/TABLES/1.

[43] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. R. Stat. Soc. Ser. B (Statistical Methodol.*, vol. 67, no. 2, pp. 301–320, Apr. 2005, doi: 10.1111/J.1467-9868.2005.00503.X.

[44] R. Tibshiranit, "Regression Shrinkage and Selection Via the Lasso," *J. R. Stat. Soc. Ser. B*, vol. 58, no. 1, pp. 267–288, Jan. 1996, doi: 10.1111/J.2517-6161.1996.TB02080.X.

[45] S. K. Reilly *et al.*, "Evolutionary changes in promoter and enhancer activity during human corticogenesis," *Science (80-.).*, vol. 347, no. 6226, pp. 1155–1159, Mar. 2015, doi: 10.1126/SCIENCE.1260943/SUPPL_FILE/TABLES5.XLSX.

[46] G. M. Cooper, E. A. Stone, G. Asimenos, E. D. Green, S. Batzoglou, and A. Sidow, "Distribution and intensity of constraint in mammalian genomic sequence," *Genome Res.*, vol. 15, no. 7, pp. 901–913, Jul. 2005, doi: 10.1101/GR.3577405.

[47] H. Nakahashi *et al.*, "A Genome-wide Map of CTCF Multivalency Redefines the CTCF Code," *Cell Rep.*, vol. 3, no. 5, pp. 1678–1689, May 2013, doi: 10.1016/J.CELREP.2013.04.024.

[48] J. C. McNally, W. G. Müller, D. Walker, R. Wolford, and G. L. Hager, "The Glucocorticoid Receptor: Rapid Exchange with Regulatory Sites in Living Cells," *Science (80-.).*, vol. 287, no. 5456, pp. 1262–1265, Feb. 2000, doi:

10.1126/SCIENCE.287.5456.1262.

[49] D. L. Moore, A. Apara, and J. L. Goldberg, “Kruppel-Like Transcription Factors in the Nervous System: Novel players in neurite outgrowth and axon regeneration,” *Mol. Cell. Neurosci.*, vol. 47, no. 4, p. 233, Aug. 2011, doi: 10.1016/J.MCN.2011.05.005.

[50] V. Lefebvre, B. Dumitriu, A. Penzo-Méndez, Y. Han, and B. Pallavi, “Control of cell fate and differentiation by Sry-related high-mobility-group box (Sox) transcription factors,” *Int. J. Biochem. Cell Biol.*, vol. 39, no. 12, pp. 2195–2214, Jan. 2007, doi: 10.1016/J.BIOCEL.2007.05.019.

[51] C. Zhao and A. Meng, “Sp1-like transcription factors are regulators of embryonic development in vertebrates,” *Dev. Growth Differ.*, vol. 47, no. 4, pp. 201–211, May 2005, doi: 10.1111/J.1440-169X.2005.00797.X.

[52] U. Borello *et al.*, “Sp8 and COUP-TF1 Reciprocally Regulate Patterning and Fgf Signaling in Cortical Progenitors,” *Cereb. Cortex*, vol. 24, no. 6, pp. 1409–1421, Jun. 2014, doi: 10.1093/CERCOR/BHS412.

[53] R. R. Waclaw *et al.*, “The Zinc Finger Transcription Factor Sp8 Regulates the Generation and Diversity of Olfactory Bulb Interneurons,” *Neuron*, vol. 49, no. 4, pp. 503–516, Feb. 2006, doi: 10.1016/J.NEURON.2006.01.018.

[54] A. W. Bruce *et al.*, “Genome-wide analysis of repressor element 1 silencing transcription factor/neuron-restrictive silencing factor (REST/NRSF) target genes,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 101, no. 28, pp. 10458–10463, Jul. 2004, doi: 10.1073/PNAS.0401827101.

[55] S. Ripke *et al.*, “Biological insights from 108 schizophrenia-associated genetic loci,” *Nat. 2014 5117510*, vol. 511, no. 7510, pp. 421–427, Jul. 2014, doi: 10.1038/nature13595.

[56] T. Bourgeron, “From the genetic architecture to synaptic plasticity in autism spectrum disorder,” *Nat. Rev. Neurosci. 2015 169*, vol. 16, no. 9, pp. 551–563, Aug. 2015, doi: 10.1038/nrn3992.

[57] S. Sood, C. M. Webera, H. C. Hodges, A. Krokhotin, A. Shalizi, and G. R. Crabtree, “CHD8 dosage regulates transcription in pluripotency and early murine neural differentiation,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 117, no. 36, p. 22331, Sep. 2020, doi: 10.1073/PNAS.1921963117/SUPPL_FILE/PNAS.1921963117.SAPP.PDF.

[58] M. T. Weirauch *et al.*, “Evaluation of methods for modeling transcription factor sequence specificity,” *Nat. Biotechnol. 2013 312*, vol. 31, no. 2, pp. 126–134, Jan. 2013, doi: 10.1038/nbt.2486.

[59] B. Statistics *et al.*, “Bayesian models for sparse regression analysis of high dimensional data.”

[60] D. J. Schaid, W. Chen, and N. B. Larson, “From genome-wide associations to candidate

causal variants by statistical fine-mapping," *Nat. Rev. Genet. 2018 198*, vol. 19, no. 8, pp. 491–504, May 2018, doi: 10.1038/s41576-018-0016-z.

[61] A. Melnikov *et al.*, "Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay," *Nat. Biotechnol. 2012 303*, vol. 30, no. 3, pp. 271–277, Feb. 2012, doi: 10.1038/nbt.2137.

[62] D. Wang *et al.*, "Comprehensive functional genomic resource and integrative model for the human brain," *Science*, vol. 362, no. 6420, Dec. 2018, doi: 10.1126/SCIENCE.AAT8464.

[63] J. Shi *et al.*, "Common variants on chromosome 6p22.1 are associated with schizophrenia," *Nat. 2009 4607256*, vol. 460, no. 7256, pp. 753–757, Jul. 2009, doi: 10.1038/nature08192.

[64] S. Zhang *et al.*, "Allele-specific open chromatin in human iPSC neurons elucidates functional disease variants," *Science (80-.).*, vol. 369, no. 6503, pp. 561–565, 2020, doi: 10.1126/science.aay3983.

[65] A. McKenna *et al.*, "The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data," *Genome Res.*, vol. 20, no. 9, p. 1297, Sep. 2010, doi: 10.1101/GR.107524.110.

[66] B. Van De Geijn, G. Mcvicker, Y. Gilad, and J. K. Pritchard, "WASP: allele-specific software for robust molecular quantitative trait locus discovery," *Nat. Methods 2015 1211*, vol. 12, no. 11, pp. 1061–1063, Sep. 2015, doi: 10.1038/nmeth.3582.

[67] J. D. Storey, "A direct approach to false discovery rates," *J. R. Stat. Soc. Ser. B (Statistical Methodol.*, vol. 64, no. 3, pp. 479–498, Aug. 2002, doi: 10.1111/1467-9868.00346.

[68] Roadmap Epigenomics Consortium *et al.*, "Integrative analysis of 111 reference human epigenomes," *Nature*, vol. 518, no. 7539, pp. 317–329, 2015, doi: 10.1038/nature14248.

[69] G. D. Smith and G. Hemani, "Mendelian randomization: genetic anchors for causal inference in epidemiological studies," *Hum. Mol. Genet.*, vol. 23, no. R1, p. R89, 2014, doi: 10.1093/HMG/DDU328.

[70] S. G. Coetzee, G. A. Coetzee, and D. J. Hazelett, "motifbreakR: an R/Bioconductor package for predicting variant effects at transcription factor binding sites," *Bioinformatics*, vol. 31, no. 23, p. 3847, Jun. 2015, doi: 10.1093/BIOINFORMATICS/BTV470.

[71] X. Wen, "Molecular QTL discovery incorporating genomic annotations using Bayesian false discovery rate control," *Ann. Appl. Stat.*, vol. 10, no. 3, pp. 1619–1638, Sep. 2016, doi: 10.1214/16-AOAS952.

[72] B. Ng *et al.*, "An xQTL map integrates the genetic architecture of the human brain's transcriptome and epigenome," *Nat. Neurosci.*, vol. 20, no. 10, p. 1418, Oct. 2017, doi: 10.1038/NN.4632.

[73] M. D. Robinson and G. K. Smyth, “Moderated statistical tests for assessing differences in tag abundance,” *Bioinformatics*, vol. 23, no. 21, pp. 2881–2887, Nov. 2007, doi: 10.1093/BIOINFORMATICS/BTM453.

[74] C. Soneson and M. D. Robinson, “Bias, robustness and scalability in single-cell differential expression analysis,” *Nat. Methods 2018 154*, vol. 15, no. 4, pp. 255–261, Feb. 2018, doi: 10.1038/nmeth.4612.

[75] J. F. Degner *et al.*, “DNase I sensitivity QTLs are a major determinant of human expression variation,” *Nat. 2012 4827385*, vol. 482, no. 7385, pp. 390–394, Feb. 2012, doi: 10.1038/nature10808.

[76] R. E. Thurman *et al.*, “The accessible chromatin landscape of the human genome,” *Nat. 2012 4897414*, vol. 489, no. 7414, pp. 75–82, Sep. 2012, doi: 10.1038/nature11232.

[77] L. de la Torre-Ubieta *et al.*, “The Dynamic Landscape of Open Chromatin during Human Cortical Neurogenesis,” *Cell*, vol. 172, no. 1–2, pp. 289-304.e18, Jan. 2018, doi: 10.1016/J.CELL.2017.12.014.

[78] J. Bryois *et al.*, “Evaluation of chromatin accessibility in prefrontal cortex of individuals with schizophrenia,” *Nat. Commun. 2018 91*, vol. 9, no. 1, pp. 1–15, Aug. 2018, doi: 10.1038/s41467-018-05379-y.

[79] J. F. Fullard *et al.*, “An atlas of chromatin accessibility in the adult human brain,” *Genome Res.*, vol. 28, no. 8, pp. 1243–1252, Aug. 2018, doi: 10.1101/GR.232488.117/-/DC1.

[80] B. E. Bernstein *et al.*, “The NIH Roadmap Epigenomics Mapping Consortium,” *Nat. Biotechnol.*, vol. 28, no. 10, p. 1045, Oct. 2010, doi: 10.1038/NBT1010-1045.

[81] A. Amiri *et al.*, “Transcriptome and epigenome landscape of human cortical development modeled in brain organoids,” *Science*, vol. 362, no. 6420, Dec. 2018, doi: 10.1126/SCIENCE.AAT6720.

[82] V. Onuchic *et al.*, “Allele-specific epigenome maps reveal sequence-dependent stochastic switching at regulatory loci,” *Science*, vol. 361, no. 6409, Sep. 2018, doi: 10.1126/SCIENCE.AAR3146.

[83] J. Xu *et al.*, “Landscape of monoallelic DNA accessibility in mouse embryonic stem cells and neural progenitor cells,” *Nat. Genet. 2017 493*, vol. 49, no. 3, pp. 377–386, Jan. 2017, doi: 10.1038/ng.3769.

[84] P. Rajarajan *et al.*, “Neuron-specific signatures in the chromosomal connectome associated with schizophrenia risk,” *Science*, vol. 362, no. 6420, Dec. 2018, doi: 10.1126/SCIENCE.AAT4311.

[85] M. Song *et al.*, “Mapping cis-regulatory chromatin contacts in neural cells links neuropsychiatric disorder risk variants to target genes,” *Nat. Genet. 2019 518*, vol. 51, no. 8, pp. 1252–1262, Jul. 2019, doi: 10.1038/s41588-019-0472-1.

[86] H. Won *et al.*, "Chromosome conformation elucidates regulatory relationships in developing human brain," *Nat. 2016 5387626*, vol. 538, no. 7626, pp. 523–527, Oct. 2016, doi: 10.1038/nature19847.

[87] K. K. Marballi and A. L. Gallitano, "Immediate Early Genes Anchor a Biological Pathway of Proteins Required for Memory Formation, Long-Term Depression and Risk for Schizophrenia," *Front. Behav. Neurosci.*, vol. 12, Feb. 2018, doi: 10.3389/FNBEH.2018.00023.

[88] I. Dulubova *et al.*, "How Tlg2p/syntaxin 16 'snares' Vps45," *EMBO J.*, vol. 21, no. 14, p. 3620, Jul. 2002, doi: 10.1093/EMBOJ/CDF381.

[89] S. Zhang *et al.*, "Multiple genes in cis mediate the effects of a single chromatin accessibility variant on aberrant synaptic development and function in human neurons," *bioRxiv*, p. 2021.12.11.472229, Dec. 2021, doi: 10.1101/2021.12.11.472229.

[90] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biol.*, vol. 15, no. 12, pp. 1–21, Dec. 2014, doi: 10.1186/S13059-014-0550-8/FIGURES/9.

[91] G. Finak *et al.*, "MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data," *Genome Biol.*, vol. 16, no. 1, pp. 1–13, Dec. 2015, doi: 10.1186/S13059-015-0844-5/FIGURES/6.

[92] F. Aguet *et al.*, "The GTEx Consortium atlas of genetic regulatory effects across human tissues," *Science*, vol. 369, no. 6509, p. 1318, Sep. 2020, doi: 10.1126/SCIENCE.AAZ1776.

[93] C. A. Vallejos, D. Risso, A. Scialdone, S. Dudoit, and J. C. Marioni, "Normalizing single-cell RNA sequencing data: challenges and opportunities," *Nat. Methods 2017 146*, vol. 14, no. 6, pp. 565–571, May 2017, doi: 10.1038/nmeth.4292.

[94] G. Chen, B. Ning, and T. Shi, "Single-cell RNA-seq technologies and related computational data analysis," *Front. Genet.*, vol. 10, no. APR, p. 317, 2019, doi: 10.3389/FGENE.2019.00317/BIBTEX.

[95] E. Shifrut *et al.*, "Genome-wide CRISPR screens in primary human T Cells reveal key regulators of immune function," *Cell*, vol. 175, no. 7, pp. 1958–1971, Dec. 2018, doi: 10.1016/J.CELL.2018.10.024.

[96] D. A. Jaitin *et al.*, "Dissecting immune circuits by linking CRISPR-pooled screens with single-cell RNA-seq," *Cell*, vol. 167, no. 7, pp. 1883--1896.e15, Dec. 2016, doi: 10.1016/J.CELL.2016.11.039.

[97] G. L. Stein-O'Brien *et al.*, "Enter the matrix: factorization uncovers knowledge from omics," *Trends Genet.*, vol. 34, no. 10, pp. 790–805, Oct. 2018, doi: 10.1016/J.TIG.2018.07.003/ATTACHMENT/43F1B901-685D-4EEB-93C9-0FC3EE022685/MMC1.DOCX.

[98] C. Eckart and G. Young, “The approximation of one matrix by another of lower rank,” *Psychom. 1936 13*, vol. 1, no. 3, pp. 211–218, Sep. 1936, doi: 10.1007/BF02288367.

[99] Z. Yang and G. Michailidis, “A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data,” *Bioinformatics*, vol. 32, no. 1, pp. 1–8, Jan. 2016, doi: 10.1093/BIOINFORMATICS/BTV544.

[100] X. Zhu, T. Ching, X. Pan, S. M. Weissman, and L. Garmire, “Detecting heterogeneity in single-cell RNA-Seq data by non-negative matrix factorization,” *PeerJ*, vol. 5, no. e2888, 2017, doi: 10.7717/PEERJ.2888/SUPP-7.

[101] S. Sra and I. Dhillon, “Generalized Nonnegative Matrix Approximations with Bregman Divergences,” in *Advances in Neural Information Processing Systems*, 2005, vol. 18, [Online]. Available: https://proceedings.neurips.cc/paper/2005/file/d58e2f077670f4de9cd7963c857f2534-Paper.pdf.

[102] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” *Nat. 1999 4016755*, vol. 401, no. 6755, pp. 788–791, Oct. 1999, doi: 10.1038/44565.

[103] D. B. Rubin and D. T. Thayer, “EM algorithms for ML factor analysis,” *Psychom. 1982 471*, vol. 47, no. 1, pp. 69–76, Mar. 1982, doi: 10.1007/BF02293851.

[104] H. Zou, T. Hastie, and R. Tibshirani, “Sparse Principal Component Analysis,” *https://doi.org/10.1198/106186006X113430*, vol. 15, no. 2, pp. 265–286, Jun. 2012, doi: 10.1198/106186006X113430.

[105] M. Lee, H. Shen, J. Z. Huang, and J. S. Marron, “Biclustering via Sparse Singular Value Decomposition,” *Biometrics*, vol. 66, no. 4, pp. 1087–1095, Dec. 2010, doi: 10.1111/J.1541-0420.2010.01392.X.

[106] L. Taslaman and B. Nilsson, “A Framework for Regularized Non-Negative Matrix Factorization, with Application to the Analysis of Gene Expression Data,” *PLoS One*, vol. 7, no. 11, p. 46331, Nov. 2012, doi: 10.1371/JOURNAL.PONE.0046331.

[107] W. Mao, E. Zaslavsky, B. M. Hartmann, S. C. Sealfon, and M. Chikina, “Pathway-level information extractor (PLIER) for gene expression data,” *Nat. Methods*, vol. 16, no. 7, pp. 607–610, Jun. 2019, doi: 10.1038/s41592-019-0456-1.

[108] S. Arlot and A. Celisse, “A survey of cross-validation procedures for model selection,” *https://doi.org/10.1214/09-SS054*, vol. 4, no. none, pp. 40–79, Jan. 2010, doi: 10.1214/09-SS054.

[109] T. Park and G. Casella, “The Bayesian Lasso,” *https://doi.org/10.1198/016214508000000337*, vol. 103, no. 482, pp. 681–686, Jun. 2012, doi: 10.1198/016214508000000337.

[110] Q. Li and N. Liny, "The Bayesian elastic net," *https://doi.org/10.1214/10-BA506*, vol. 5, no. 1, pp. 151–170, Mar. 2010, doi: 10.1214/10-BA506.

[111] A. E. Hoerl and R. W. Kennard, "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970, doi: 10.1080/00401706.1970.10488634.

[112] B. E. Engelhardt and M. Stephens, "Analysis of Population Structure: A Unifying Framework and Novel Methods Based on Sparse Factor Analysis," *PLoS Genet.*, vol. 6, no. 9, Sep. 2010, doi: 10.1371/JOURNAL.PGEN.1001117.

[113] T. J. Mitchell and J. J. Beauchamp, "Bayesian Variable Selection in Linear Regression," *J. Am. Stat. Assoc.*, vol. 83, no. 404, p. 1023, Dec. 1988, doi: 10.2307/2290129.

[114] C. M. Carvalho, J. Chang, J. E. Lucas, J. R. Nevins, Q. Wang, and M. West, "High-dimensional sparse factor modeling: applications in gene expression genomics," *J. Am. Stat. Assoc.*, vol. 103, no. 484, pp. 1438–1456, Dec. 2008, doi: 10.1198/016214508000000869.

[115] K. Bunte, E. Leppäaho, I. Saarinen, and S. Kaski, "Sparse group factor analysis for biclustering of multiple data sources," *Bioinformatics*, vol. 32, no. 16, pp. 2457–2463, Aug. 2016, doi: 10.1093/BIOINFORMATICS/BTW207.

[116] E. I. George and R. E. McCulloch, "Variable selection via Gibbs sampling," *J. Am. Stat. Assoc.*, vol. 88, no. 423, pp. 881–889, Sep. 1993, doi: 10.2307/2290777.

[117] M. Stephens, "False discovery rates: A new deal," *Biostatistics*, vol. 18, no. 2, pp. 275–294, Apr. 2017, doi: 10.1093/biostatistics/kxw041.

[118] W. Wang and M. Stephens, "Empirical Bayes Matrix Factorization," *J. Mach. Learn. Res.*, vol. 22, pp. 1–40, 2021, Accessed: Mar. 20, 2022. [Online]. Available: http://jmlr.org/papers/v22/20-589.html.

[119] M. A. van de Wiel, D. E. Te Beest, and M. M. Münch, "Learning from a lot: Empirical Bayes for high-dimensional model-based prediction," *Scand. Stat. Theory Appl.*, vol. 46, no. 1, p. 2, Mar. 2019, doi: 10.1111/SJOS.12335.

[120] C. Andrieu, N. De Freitas, A. Doucet, and M. I. Jordan, "An Introduction to MCMC for Machine Learning," *Mach. Learn. 2003 501*, vol. 50, no. 1, pp. 5–43, Jan. 2003, doi: 10.1023/A:1020281327116.

[121] S. van Erp, D. L. Oberski, and J. Mulder, "Shrinkage priors for Bayesian penalized regression," *J. Math. Psychol.*, vol. 89, pp. 31–50, Apr. 2019, doi: 10.1016/J.JMP.2018.12.004.

[122] D. Knowles and Z. Ghahramani, "Nonparametric Bayesian sparse factor models with application to gene expression modeling," *Ann. Appl. Stat.*, vol. 5, no. 2B, pp. 1534–1552, Jun. 2011, doi: 10.1214/10-AOAS435.

[123] A. Bhattacharya and D. B. Dunson, “Sparse Bayesian infinite factor models,” *Biometrika*, vol. 98, no. 2, p. 291, Jun. 2011, doi: 10.1093/BIOMET/ASR013.

[124] R. B. O’Hara and M. J. Sillanpää, “A review of bayesian variable selection methods: What, how and which,” *Bayesian Anal.*, vol. 4, no. 1, pp. 85–118, 2009, doi: 10.1214/09-BA403.

[125] Y. Guan and M. Stephens, “Bayesian variable selection regression for genome-wide association studies and other large-scale problems,” *Ann. Appl. Stat.*, vol. 5, no. 3, pp. 1780–1815, Sep. 2011, doi: 10.1214/11-AOAS455.

[126] B. L. WELCH, “The generalisation of student’s problems when several different population variances are involved,” *Biometrika*, vol. 34, no. 1–2, pp. 28–35, 1947, doi: 10.1093/BIOMET/34.1-2.28.

[127] C. J. Sherr and J. M. Roberts, “CDK inhibitors: positive and negative regulators of G1-phase progression,” *Genes Dev.*, vol. 13, no. 12, pp. 1501–1512, Jun. 1999, doi: 10.1101/GAD.13.12.1501.

[128] N. P. D. Liau *et al.*, “The molecular basis of JAK/STAT inhibition by SOCS1,” *Nat. Commun. 2018 91*, vol. 9, no. 1, pp. 1–14, Apr. 2018, doi: 10.1038/s41467-018-04013-1.

[129] T. Aso, A. Shilatifard, J. W. Conaway, and R. C. Conaway, “Transcription syndromes and the role of RNA polymerase II general transcription factors in human disease.,” *J. Clin. Invest.*, vol. 97, no. 7, p. 1561, Apr. 1996, doi: 10.1172/JCI118580.

[130] D. C. Palmer and N. P. Restifo, “Suppressors of Cytokine Signaling (SOCS) in T cell differentiation, maturation, and function,” *Trends Immunol.*, vol. 30, no. 12, p. 592, Dec. 2009, doi: 10.1016/J.IT.2009.09.009.

[131] J. Huang, Y. L. Zhao, Y. Li, J. A. Fletcher, and S. Xiao, “Genomic and functional evidence for an ARID1A tumor suppressor role,” *Genes, Chromosom. Cancer*, vol. 46, no. 8, pp. 745–750, Aug. 2007, doi: 10.1002/GCC.20459.

[132] S. Jones *et al.*, “Somatic mutations in the chromatin remodeling gene ARID1A occur in several tumor types,” *Hum. Mutat.*, vol. 33, no. 1, pp. 100–103, Jan. 2012, doi: 10.1002/HUMU.21633.

[133] R. C. Wu, T. L. Wang, and I. M. Shih, “The emerging roles of ARID1A in tumor suppression,” *Cancer Biol. Ther.*, vol. 15, no. 6, pp. 655–664, 2014, doi: 10.4161/CBT.28411.

[134] J. Li *et al.*, “Epigenetic driver mutations in ARID1A shape cancer immune phenotype and immunotherapy,” *J. Clin. Invest.*, vol. 130, no. 5, pp. 2712–2726, May 2020, doi: 10.1172/JCI134402.

[135] M. A. Lalli, D. Avey, J. D. Dougherty, J. Milbrandt, and R. D. Mitra, “High-throughput single-cell functional elucidation of neurodevelopmental disease-associated genes reveals

convergent mechanisms altering neuronal differentiation," *Genome Res.*, vol. 30, no. 9, pp. 1317–1331, Oct. 2020, doi: 10.1101/GR.262295.120/-/DC1.

[136] A. Sessa *et al.*, "SETD5 regulates chromatin methylation state and preserves global transcriptional fidelity during brain development and neuronal wiring," *Neuron*, vol. 104, no. 2, pp. 271--289.e13, Oct. 2019, doi: 10.1016/J.NEURON.2019.07.013.

[137] F. Aguet *et al.*, "Genetic effects on gene expression across human tissues," *Nature*, vol. 550, no. 7675, pp. 204–213, Oct. 2017, doi: 10.1038/nature24277.

[138] W. J. Astle *et al.*, "The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease," *Cell*, vol. 167, no. 5, pp. 1415-1429.e19, Nov. 2016, doi: 10.1016/J.CELL.2016.10.042/ATTACHMENT/71619ACE-8DF9-44BD-AA85-84005E16B095/MMC7.XLSX.

[139] A. T. L. Lun, K. Bach, and J. C. Marioni, "Pooling across cells to normalize single-cell RNA sequencing data with many zero counts," *Genome Biol.*, vol. 17, no. 1, pp. 1–14, Apr. 2016, doi: 10.1186/S13059-016-0947-7/TABLES/2.

[140] A. Lun, "Overcoming systematic errors caused by log-transformation of normalized single-cell RNA sequencing data," *bioRxiv*, p. 404962, Aug. 2018, doi: 10.1101/404962.

[141] R. Satija, J. A. Farrell, D. Gennert, A. F. Schier, and A. Regev, "Spatial reconstruction of single-cell gene expression data," *Nat. Biotechnol.*, vol. 33, no. 5, pp. 495–502, Apr. 2015, doi: 10.1038/nbt.3192.

[142] J. Wang, S. Vasaikar, Z. Shi, M. Greer, and B. Zhang, "WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit," *Nucleic Acids Res.*, vol. 45, no. W1, pp. W130–W137, Jul. 2017, doi: 10.1093/NAR/GKX356.

[143] A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo, and J. P. Mesirov, "Molecular signatures database (MSigDB) 3.0," *Bioinformatics*, vol. 27, no. 12, pp. 1739–1740, Jun. 2011, doi: 10.1093/BIOINFORMATICS/BTR260.

[144] F. Buettner, N. Pratanwanich, D. J. McCarthy, J. C. Marioni, and O. Stegle, "f-scLVM: scalable and versatile factor analysis for single-cell RNA-seq," *Genome Biol.*, vol. 18, no. 1, p. 218, Nov. 2017, doi: 10.1186/S13059-017-1334-8.

[145] R. Argelaguet *et al.*, "MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data," *Genome Biol.*, vol. 21, no. 1, p. 111, May 2020, doi: 10.1186/S13059-020-02015-1.

[146] B. Duan *et al.*, "Model-based understanding of single-cell CRISPR screening," *Nat. Commun.*, vol. 10, no. 1, pp. 1–11, May 2019, doi: 10.1038/s41467-019-10216-x.

[147] J. Fan, Y. Liao, and W. Wang, "Projected principal component analysis in factor models," *Ann. Stat.*, vol. 44, no. 1, pp. 219–254, Feb. 2016, doi: 10.1214/15-AOS1364.

[148] G. Li, D. Yang, A. B. Nobel, and H. Shen, "Supervised singular value decomposition and its asymptotic properties," *J. Multivar. Anal.*, vol. 146, pp. 7–17, Apr. 2016, doi: 10.1016/J.JMVA.2015.02.016.

[149] S. Yu, K. Yu, V. Tresp, H.-P. Kriegel, and M. Wu, "Supervised probabilistic principal component analysis," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006, pp. 464–473.

[150] S. Zamani Dadaneh, M. Zhou, and X. Qian, "Covariate-dependent negative binomial factor analysis of RNA sequencing data," *Bioinformatics*, vol. 34, no. 13, pp. i61--i69, Jul. 2018, doi: 10.1093/BIOINFORMATICS/BTY237.

[151] M. Levo and E. Segal, "In pursuit of design principles of regulatory sequences," *Nat. Rev. Genet.*, vol. 15, no. 7, pp. 453–468, 2014, doi: 10.1038/nrg3684.

[152] N. Kumasaka, A. J. Knights, and D. J. Gaffney, "Fine-mapping cellular QTLs with RASQUAL and ATAC-seq," *Nat. Genet. 2015 482*, vol. 48, no. 2, pp. 206–213, Dec. 2015, doi: 10.1038/ng.3467.

[153] Z. Zhang *et al.*, "Genetic analyses support the contribution of mRNA N6-methyladenosine (m6A) modification to human disease heritability," *Nat. Genet. 2020 529*, vol. 52, no. 9, pp. 939–949, Jun. 2020, doi: 10.1038/s41588-020-0644-z.

[154] S. Xie, J. Duan, B. Li, P. Zhou, and G. C. Hon, "Multiplexed Engineering and Analysis of Combinatorial Enhancer Activity in Single Cells," *Mol. Cell*, vol. 66, no. 2, pp. 285-299.e5, Apr. 2017, doi: 10.1016/j.molcel.2017.03.007.

[155] X. H. Zhang, L. Y. Tee, X. G. Wang, Q. S. Huang, and S. H. Yang, "Off-target Effects in CRISPR/Cas9-mediated Genome Engineering," *Mol. Ther. - Nucleic Acids*, vol. 4, no. 11, p. e264, Jan. 2015, doi: 10.1038/MTNA.2015.37.

[156] J. G. Doench *et al.*, "Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9," *Nat. Biotechnol. 2015 342*, vol. 34, no. 2, pp. 184–191, Jan. 2016, doi: 10.1038/nbt.3437.

[157] L. Wang, "Single-cell normalization and association testing unifying CRISPR screen and gene co-expression analyses with Normalisr," *Nat. Commun. 2021 121*, vol. 12, no. 1, pp. 1–13, Nov. 2021, doi: 10.1038/s41467-021-26682-1.

[158] M. Gasperini *et al.*, "CRISPR/Cas9-Mediated Scanning for Regulatory Elements Required for HPRT1 Expression via Thousands of Large, Programmed Genomic Deletions," *Am. J. Hum. Genet.*, vol. 101, no. 2, pp. 192–205, Aug. 2017, doi: 10.1016/J.AJHG.2017.06.010.

[159] Y. Diao *et al.*, "A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells," *Nat. Methods*, vol. 14, no. 6, pp. 629–635, May 2017, doi: 10.1038/NMETH.4264.

[160] S. E. Pierce, J. M. Granja, and W. J. Greenleaf, "High-throughput single-cell chromatin

accessibility CRISPR screens enable unbiased identification of regulatory networks in cancer," *Nat. Commun. 2021 121*, vol. 12, no. 1, pp. 1–8, May 2021, doi: 10.1038/s41467-021-23213-w.

[161] N. Liscovitch-Brauer *et al.*, "Profiling the genetic determinants of chromatin accessibility with scalable single-cell CRISPR screens," *Nat. Biotechnol. 2021 3910*, vol. 39, no. 10, pp. 1270–1277, Apr. 2021, doi: 10.1038/s41587-021-00902-x.

[162] E. P. Mimitou *et al.*, "Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells," *Nat. Methods 2019 165*, vol. 16, no. 5, pp. 409–412, Apr. 2019, doi: 10.1038/s41592-019-0392-0.